DE MONTFORT UNIVERSITY, LEICESTER

The development of Deoxyribonucleic Acid (DNA) based methods for the identification and authentication of medicinal plant material.

Caroline Howard

Ph.D.

June 2010

The development of Deoxyribonucleic Acid (DNA) based methods for the identification and authentication of medicinal plant material.

Caroline Howard

A thesis submitted in partial fulfilment

of the requirement for the

degree of Doctor of Philosophy

to De Montfort University

Faculty of Health and Life Sciences

June 2010

Abstract

Herbal medicines are growing in popularity in the Western world and are becoming more stringently regulated under new EU legislation. Within the arena of herbal medicines, St. John's Wort (SJW), *Hypericum perforatum,* is a top ten best seller with clinical evidence to support its use as an anti-depressant.

A fundamental requirement of the new legislation is to prove the identity of the plant material in question. This is currently achieved via morphological and chemical methods, neither of which are ideal. A wide range of DNA based methods have been applied to this arena, standardisation is required to realise the potential of DNA based techniques.

The DNA barcoding initiative aims to produce sequence data for all plant species, capable of species identification. The proposal is to use these data to design fast and effective DNA based methods of identification. For assay design, the putative barcode region nrITS was selected as a platform.

Three assays were designed;

- A PCR assay designed to hyper variable sequences within a barcode region. This assay is capable of distinguishing SJW from other closely related species.
- A quantitative qPCR assay designed to measure total DNA and specific SJW DNA within a mixed sample.
- A multiplex PCR incorporating fluorescently labelled primers, allowing amplicon detection by capillary electrophoresis. This assay identifies four separate *Hypericum* species, including SJW, with a mixed sample in one reaction.

The suitability of the nrITS and three other barcode regions is assessed based on sequence data generated for 32 vouchered samples of different *Hypericum* species, and a Lithuanian sample set of 22 and 16 *H. perforatum* and *H. maculatum* samples respectively. The *matK* is currently unusable, the *rbcL* highly conserved, *trnH-psbA* problematically variable and the nrITS proved to be ideal for assay design.

Acknowledgements

I would like to thank my entire supervisory team for initiating this project, seeking funding and supporting every aspect of my research. In particular, Dr Adrian Slater for conceiving the project, and for guiding and advising with infinite patience, this was tested. Dr Paul Bremner for his optimistic and global approach, and Dr Mark Fowler for his insightful critique and advice.

Thanks are due to the many contributors of samples to this research. Mark Carine at the Natural History Museum London for providing vouchered DNA samples and general guidance. Asta Judzentiene of the Institute of Chemistry Vilnius, Lithuania, for providing *Hypericum* samples. Santiago Madriñán, Univ. de los Andes, for permission to use DNA from vouchered specimens which was supplied by the Kew DNA Databank.

I would also like to thank our collaborators from Leicester University, Dr Eleanor Graham and Eleni Socratous for their work and expertise. Thanks are also due to my colleagues in the Cell Signalling Laboratory of De Montfort University where I have conducted this research.

I would also like to thank our Research Assistant, Sarah Smith, who worked on this project both with and without payment. Her hard work was and is very much appreciated.

Finally I would like to thank my husband Alex, who has been patient and supportive throughout my research, and all of my friends and family.

Publications and Presentations

Original Papers

Molecular Identification of Hypericum perforatum L. by PCR amplification of the ITS and 5.8S rDNA Region. Caroline Howard, Paul D. Bremner, Mark R. Fowler, Belinda Isodo, Nigel W. Scott and Adrian Slater. Planta Medica 2009: 75; 8, 864-869.

Primer Design for DNA Based Identification of Medicinal Plants. C. Howard, P.D. Bremner, M.R. Fowler, N.W. Scott and A. Slater. Acta Horticulturae: In Press.

Presentations

PlantID – A System for the Identification of Medicinal Plant Material by DNA Profiling. Caroline Howard, Eleni Socratous, Sarah Smith, Eleanor Graham, Mark R Fowler, Nigel W Scott, Paul D Bremner and Adrian Slater. RRJ Arroo (ed.) (2010) Trends in Natural Products Research: Abstracts of the Phytochemical Society of Europe. Phytochemical Society of Europe, Leicester. ISBN 978-0-9565472-0-0. Meeting held at De Montfort University Leicester, April 2010 .

Development of DNA sequence-based tests to discriminate Hypericum species. Caroline Howard, Mark Fowler, Paul Bremner, Nigel Scott and Adrian Slater. Invited talk given at the 2nd Annual Hypericum Meeting at the IPK in Gatersleben Germany, November 2009.

Molecular Identification of Hypericum perforatum L. by PCR amplification of the rDNA ITS region. Caroline Howard, Paul Bremner, Mark Fowler and Adrian Slater. The African Journal of Traditional Complementary and Alternative Medicine 2009; World Conference on Medicinal and Aromatic Plants, Cape Town South Africa, November 2008.

Molecular Identification of St John's Wort by PCR amplification of the ITS1 Region: Implications for medicinal plant identification. Caroline Howard, Paul D. Bremner, Mark R. Fowler and Adrian Slater. Journal of Pharmacy and Pharmacology 2008; 60: A10. British Pharmaceutical Conference, September 2008.

DNA based identification of *Hypericum perforatum* /Medicinal Plants. Caroline Howard. Oral presentation to a meeting of King's College London, Royal Botanic Gardens Kew and The London School of Pharmacy, February 2008.

Posters

Molecular identification and quantitation of Hypericum perforatum in mixed samples. Caroline Howard, Paul D Bremner, Mark R Fowler, Nigel W Scott and Adrian Slater. Planta Medica

2009: 75; 9, 1000. 55th International Congress and Annual Meeting of the Society for Medicinal Plant Research and Natural Product Research, Geneva 2009.

Contents

Figures

Tables

1 Introduction

1.1 Medicinal Plant Usage and Regulation

1.1.1 Medicinal Plant Usage

Mankind's use of plants for their medicinal properties can be traced back to the origins of the written word. The Ancient Egyptians produced papyrus scripts recording the use of plants such as aloe, thyme, juniper, cannabis and opium in medicine (Aboelsoud, 2010). Over thousands of years, the knowledge of the application and administration of plants in the prevention and treatment of human disease and illness has grown and developed into sophisticated holistic lifestyle systems, such as Ayurveda and Traditional Chinese Medicine (TCM). These systems are widely used, and traditional medicine remains a primary source of health care for up to 80% of the population of some African and Asian countries (World Health Organization, 2010), whether due to individual choice or economic and other constraints.

Until the twentieth century, plants were the main source of therapeuticals (Marston and Hostettmann, 2009), though they became largely overlooked in More Economically Developed Countries (MEDCs) with the advent of modern pharmaceuticals, despite many of these proprietary drugs being originally developed from plant derived compounds. Classic examples of this include: aspirin (acetylsalicylic acid) formerly derived from the bark of the willow tree, (*Salix spp*.) and used as an analgesic; digitalin, isolated from the foxglove (*Digitalis spp*.) used in the treatment of irregular heart rhythms; and atropine, found in deadly nightshade (*Atropa belladonna)* amongst other plants, and used for purposes including pupil dilation.

More recently, there has been a trend back toward medicinal plant usage in the UK and other MEDCs, this is usually referred to under the category of 'Complementary and Alternative Medicine' (CAM). This term encompasses a large range of treatments and therapies, from homeopathic remedies to essential oils, acupuncture and massage to herbal preparations. CAMs are gaining in popularity in MEDCs, and between 2002 and 2007 in the USA use by adults grew from 36 to 39.3% (National Institutes of Health, 2008). In the UK alone CAM products were worth an estimated £115 million in 2000, which represents a growth of 23% on 1998 figures (Barnes, 2003).

Herbal medicines in MEDCs have developed into a large industry. Medicinal plants are one of the most economically valuable and profitable areas of CAM, generating billions of dollars in

revenue (World Health Organization, 2010). A survey carried out in 2008 found that 35% of UK adults have used herbal medicines (Medicines and Healthcare products Regulatory Agency, 2009) and another in the US found that they had been used by 17.7% of adults in 2007 (National Institutes of Health, 2008). In Germany, where herbal medicines are widely used and accepted as efficacious, they are termed 'phytomedicines' or 'phytotherapeutic agents' and are prescribed in an evidence-based approach according to documented clinical value (Barnes, 2003).

The recent increase in usage of medicinal plant products has highlighted the necessity for assurances of safety and quality in commercial products. These have long been questioned due to 'continuing evidence of an international trade in herbal remedies made to an unreliable standard' (UK House of Commons, 2003). Recent high profile cases have emphasised the potential for harm. A woman prescribed pills by a TCM practitioner containing the dangerous (and currently banned) substance, aristolochic acid, was diagnosed with kidney failure followed by cancer of the urinary tract. These illnesses were both attributed to the pills, which were administered at a much higher dosage than would have been recommended in TCM (http://news.bbc.co.uk/1/low/health/8520171.stm).

This substance had been brought to the attention of regulators previously when slimming products were found to contain *Aristolochia* plant material, and consequently aristolochic acid. This resulted in 100 cases of renal failure, (Vanherweghem, 1998), and calls for more stringent regulation of the industry. This is particularly necessary when 58% of UK adults who use herbal medicines agree with the statement that 'herbal medicines are safe because they are natural' (Medicines and Healthcare products Regulatory Agency, 2009).

1.1.2 Medicinal Plant Legislation

Medicinal plant products for human use in the European Union (EU) are regulated by the Traditional Herbal Medicines Directive (Directive 2004/24/EC). This recently introduced legislation requires that the medicinal plant in question can be shown to have been used traditionally in the EU for 30 years, or for 15 within the EU and 15 years elsewhere making a total of 30 years for plants from outside the EU (Vlietinck et al., 2009). It also requires that safety data can be provided, though this can be literature based, and that the producer can guarantee the quality of their product with reference to Good Manufacturing Process (GMP) (Barnes, 2003). Another strategy referenced is The World Health Organisation guidelines on Good Agricultural and Collection Practices for Medicinal Plants which spell out requirements

for species identification, collection practices and cultivation of medicinal plant species (http://whqlibdoc.who.int/publications/2003/9241546271.pdf). The claims on the labels of medicinal plant products also fall under this legislation, and will become standardised as to what claims may be made about the product and the safety information which must be included.

The main aim of the Traditional Herbal Medicines Directive is to regulate and control herbal medicines under an EU wide mandate, providing quality and safety assurance.

1.1.2.1 Definition of a Medicinal Plant Product

A complicated boundary exists between herbal medicines and food supplements, which are regulated by the Food Standards Authority under The Food Supplements Directive (2002/46/EC). The MHRA has a dedicated 'Borderline' section tasked with allocating products as either traditional herbal medicines or food supplements.

Traditional Herbal Medicines (THMs) and Food Supplements are defined as follows:

THMs

"A product presented for treating or preventing disease, or which may be administered with a view to restoring, correcting or modifying physiological function in humans, falls within the definition of a medicinal product and is subject to the requirements of the Medicines Directive"

Food Supplements

"food supplement means any food the purpose of which is to supplement the

normal diet and which –

a) is a concentrated source of a vitamin or mineral or other substance with a nutritional or physiological effect, alone or in combination; and

b) is sold in dose form"

(Food Standards Agency, 2005)

The main distinction made between these two definitions is the treatment and prevention of disease, which may only be claimed by medicinal products.

1.1.2.2 Legislation Requirements

The directive can be complied with in two ways; medicinal plant products can gain either a "Marketing Authorisation" or "Registration" under the Traditional Herbal Medicines Registration Scheme (THMRS). Marketing Authorisations require, amongst other measures,

that the product be proven to be efficacious. This can either be shown in new clinical trials for safety and efficacy, or via a 'bibliographic application' in which the active compound of the medicinal plant in question can be shown to have been well established in the EU for 10 years (Vlietinck et al., 2009).

This is not possible when the active compounds of a medicinal plant are unknown, or not fully understood. In this situation a THM Registration Application may be obtained. This requires that the efficacy is plausible based on the long term traditional use of the plant, as described in section 1.1.2 (Vlietinck et al., 2009).

This directive came in to force in 2004, though products already on the market are required to comply by 2011. This will restrict the number and variety of medicinal plant products on sale, as either process, marketing authorisation or registration, is lengthy and expensive. It will also increase the quality and efficacy of products, and may aid manufacturers in increasing public confidence in their products, and herbal medicines in general.

One of the key requirements of both Registration and Market Authorisation is 'Quality aspects', referring to the need for identification and authentication of plant material upstream of manufacturing and processing. Incorrect identification of plant material can lead to inadvertent contamination or even replacement of the target species. It is also the case that some of the more rare and expensive medicinal plant species are intentionally substituted for less valuable alternatives (Joshi et al., 2004). Reliable identification methods are therefore paramount in assuring the quality and safety of herbal medicinal products.

1.2 Current Medicinal Plant Identification Methods

The European Pharmacopoeia was created in 1964 after European member states signed the Convention on the Elaboration of a European Pharmacopoeia, which would supersede those of the individual member states. The Pharmacopoeias are reference publications produced under the European Directorate for the Quality of Medicines and HealthCare. It sets out the mandatory requirements in Europe for medicines for human or veterinary use, detailing and quantifying the essential components and how they should be measured. These standards are decided upon by panels of experts, each of which is chaired by a member of the European Pharmacopoeia Commission. Each separate medicine has a dedicated monograph, detailing its composition, use, testing procedures and specific properties (European Directorate for the Quality of Medicines and HealthCare, 2010). This includes medicinal plants, each species is

described in a monograph detailing morphological features along with the agreed compositional requirements and how these should be measured.

1.2.1 Morphology

Morphological methods of identification fall into two main categories; macroscopic and microscopic. Macroscopic refers to the visualisation of all or a large section of the plant in order to compare physical attributes. Attributes such as shape, size, colour, texture and odour of leaves, flowers and fruits are compared directly to a reference sample (Joshi et al., 2004). This requires a highly trained individual and a large and current reference library to make the identification (Techen et al., 2004). Also, in some cases the entire plant is required, and in many cases the plant must be flowering. Microscopic identification overcomes some of these problems, as the physical attributes required for identification are much smaller. However, limitations with this method are access to reference samples and the stage in the plant life cycle of the sample, since plant age can also affect morphology.

Highly qualified individuals are required for both macroscopic and microscopic identification. The number of trained taxonomists has been reducing in recent years, and a lack of interest in the area from the current generation of students has put this skill at a premium (Smith and Figueiredo, 2009). This promises to reduce numbers further, and cause the approach to become inaccessible to potential users.

Morphological identification must be carried out in line with the European Pharmacopoeial monographs in order to satisfy the Quality criteria of the European Directive for medicinal plants for human use. If there is not a monograph available for the plant in question, a detailed description must be put in to place that meets the requirements for monograph production (Vlietinck et al., 2009).

1.2.2 Chemical Methods

In most instances, the identity of plant material used in medicinal products is confirmed by chemical analysis of the compounds present. These are usually in one of two forms, testing for biomarkers or analysing an entire 'chemical fingerprint'. The requirements for each plant species are also laid out in the Pharmacopoeial Monographs.

1.2.2.1 Thin Layer Chromatography (TLC)

This method separates compounds based on their movement on a solid phase when carried by a mobile phase. Compounds will travel different distances based on their chemical properties

and the affect this has on their affinity to both the mobile and solid phases. This creates a profile for all compounds present when visualised under different conditions. In some cases further reagents and different light sources are required for the visualisation of compounds. A degree of standardisation is accomplished using the retention factor, Rf. This is a measure of the distance travelled by a compound in relation to the mobile phase, or solvent. The Rf value of an unknown compound can be compared to a known standard to aid identification, this should be on the same plate as external factors affect Rf values. The application of this technique for medicinal plants is based on marker compounds, which are characterised as to colour and position for each species. The technique is low cost, simple and fast but problems include a lack of automation and reproducibility, though these are minor compared to the ease of application of the technique (Marston and Hostettmann, 2009).

1.2.2.2 High Performance Liquid Chromatography (HPLC)

Based on a similar premise to TLC, HPLC is a column based chromatography method. The mobile phase is pushed through the column by a pump, creating high pressure, and the time taken to pass through, retention time, is indicative of the compounds present. The detector used may also give information as to the identity of the compound based on spectrometric data. This method provides a means of measuring quality and quantity, when paired with an appropriate detector, and can be automated to some extent though sample preparation remains manual. This method has become the most widely used chromatographic method, and is used in many areas including pharmaceutical validation (Marston and Hostettmann, 2009).

These methods must be used in conjunction with one another to fulfil the requirements of the European Directive. TLC with HPLC is acceptable, but also HPLC coupled with different detection methods such as ultraviolet-diode array or mass spectrometry are suitable (Vlietinck et al., 2009).

Chemical analysis is the most powerful identification technique currently in use. It is very precise and requires much less input plant material than morphological methods. The restrictions of these methods are the nature and identity of the compounds measured. Medicinal plants are living organisms, and as such show variation in response to environmental triggers. The chemical constituents of a plant are affected by many factors such as the climate in which the plants were grown, the time of year they were harvested and post-harvest treatment and storage (Barnes, 2003, Joshi et al., 2004, Khan, 2006, Vlietinck et al., 2009).

Manufacturers often standardise their products in an attempt to address the variability in raw material; this is highly effective when the active compound is known. However, for many medicinal plants the active compound is not known, or is thought to be a combination of the many chemical compounds naturally occurring within a specific plant. In these situations preparations may be standardised to a 'marker compound', one chosen to be an indicator for the plant species of interest.

Refined herbal extracts have been subjected to purification procedures which increase the concentration of candidate active compounds to a desired level (Vlietinck et al., 2009). During this process it is possible that other compounds could be removed, and the proportions of compounds found within the medicinal plant being altered.

Many of the biologically active compounds measured using these methods are present in more than one species. Due to this, while confirming the presence of active or potentially active compounds, these methods do not necessarily discriminate to the species level.

In these situations, a method to identify the species of the starting plant material can be the only meaningful identification test (Joshi et al., 2004, Khan, 2006).

1.3 DNA-Based Identification of Medicinal Plant Material

An alternative to chemical and morphological identification is presented by DNA based methods. The genetic information contained within all plant material is highly complex and abundant; the challenge is to develop fast and efficient methods to utilise this information. Since this technology has been made available to researchers, many different methods have been designed for its use. Some of the main options are described in this section.

1.3.1 Randomly Amplified Polymorphic DNA (RAPD)

RAPD is a Polymerase Chain Reaction (PCR) based technique which enables the design of identification assays without prior knowledge of the DNA sequence to be targeted. This is the main advantage of RAPD, as lengthy and expensive sequencing efforts are not required.

The basis of RAPD is the use of sets of random primers typically 10-15 bases in length. These primers anneal at positions throughout the genome of the sample, and a number will anneal at a distance from, and orientation to each other which will allow the production of a PCR amplicon. The number and size of the amplicons produced are then analysed by gel

electrophoresis, where multiple bands are visualised. Sequence variations in different genomes cause the production of different sized amplicons, and therefore a different pattern of banding is created. The banding patterns produced by closely related species may then be aligned and scored, based on the presence or absence of a band at a given position (1 – presence, 0 – absence) (Smelcerovic et al., 2006). The bands which are polymorphic, and only present in one species can then be used as identifiers.

This technique has been developed to identify a number of medicinal plant species. For example, a RAPD method was used to differentiate the nine different medicinally used species of *Dendrobium*, in the Orchidaceae family (Zha et al., 2009). These species are difficult to discriminate based on morphology as they share many features, and the plant material is typically dried and processed prior to sale.

A widely used Ayurvedic preparation, Dashmoola, should be constituted of the roots of ten different plant species, one of which is *Desmodium gangeticum,* but substitution of raw materials is common. To address this, RAPD was used to create a species specific assay for *D. gangeticum*, and two potential adulterants or substitutes *D. velutinum* and *D. triflorum* (Irshad et al., 2009).

The RAPD technique was also used to analyse six *Hypericum* species from Serbia and investigate correlations between genetic patterns and data from phytochemical analysis and classical taxonomy (Smelcerovic et al., 2006). Other DNA based techniques were also tested, but the greatest correlation was found with the RAPD data.

Two genera used traditionally in the relief of menopausal symptoms, *Actaea* and *Trifolium*, have been investigated with RAPD. The findings showed that all three of the species tested from the two genera produced specific banding patterns which were not affected by the age of the plant material, though samples from different geographical areas produced slightly altered patterns (Xu et al., 2002). This was recommended as an identification technique to be used when only powdered plant material is available, particularly for the most economically important species, *A. racemosa* and *T. pratense*.

However, the RAPD method is notoriously unreliable, with different banding patterns often occurring between laboratories (Yip et al., 2007). The PCR conditions are of the upmost importance in this method, particularly the annealing temperatures in the initial cycles; any deviation can directly affect the results. Due to these restrictions, the technique is now widely

used as an upstream process for the development of other methods, as in the *Desmodium* example described (Irshad et al., 2009). Further adaptations include the use of fluorescent labels, such as in Direct Amplification of Length Polymorphism (DALP). Here two primers of different sequence are used, one forward and one reverse, each with a different fluorescent label attached. The resultant amplicons are analysed via capillary electrophoresis, enabling the size of fragments to be measured and the type of fluorescent label attached recorded. This technique has been used to differentiate between wild and cultivated *Panax ginseng* (Wang et al., 2004). The wild ginseng was found to have much more genetic variety than the cultivated resulting in different DALP fingerprints.

1.3.2 Sequence Characterised Amplified Region (SCAR)

SCAR is an adaptation of RAPD which confers stability and reliability to the technique. After RAPD analysis, specific polymorphic bands are identified which are excised from electrophoresis gels. These are then sequenced, allowing PCR primers to be designed as an exact match to the amplicon (Moon et al., 2010). In this situation, RAPD is essentially an aid to primer design. As the primers are designed to match the target sequence exactly, the PCR conditions for the final assay can be very stringent so as to assist specificity.

This technique has been developed for 'Qianhu', a traditional medicine used in the treatment of lung diseases. This is described in the Chinese pharmacopoeia as the plant species *Peucedanum praeruptorum*, but in the Japanese and Korean pharmacopoeias as a mixture of this and *Peucedanum decursivum* (also known as *Angelica decursiva*). To further complicate this situation, *Anthricus sylvestris* is cultivated and distributed under a similar name, and looks very similar to the aforementioned plant species, so is often mistaken for Qianhu. The SCAR technique was successfully used to differentiate each of the three species, and was further developed into a multiplex technique allowing each to be identified in one reaction (Choo et al., 2009). Following this, the same group developed the technique for *Cynanchum wilfordii*, *C. Auriculatum* and *Polygonum multiflorum*, each of which are used in traditional Chinese, Japanese and Korean medicine. Due to the morphological similarity of the parts used, the dried root tubers, they are often interchanged. Again, the technique was capable of distinguishing each species and developed into a multiplex assay (Moon et al., 2010).

The Ayurvedic treatment Vidarikand used for menstrual problems and pains should be made from the plant species *Pueraria tuberose.* However, at least three other species are sold under

this name. The SCAR method was used to produce an assay to discriminate the correct species from known adulterants (Devaiah and Venkatasubramanian, 2008).

1.3.3 Restriction Fragment Length Polymorphism (RFLP)

RFLP is divided into two categories, PCR-RFLP and genomic RFLP, but both methods depend on the specificity of restriction endonucleases which cleave DNA at a particular sequence. In the first method, a DNA region is amplified by PCR and then subjected to a restriction enzyme digest. Depending on the frequency of occurrence of the recognition sequence, the PCR amplicon is cleaved a number of times. This creates several differently sized products which can be analysed via gel electrophoresis in the same way as RAPD fragments. The use of multiple endonucleases increases the likelihood of cleavage, and/or the number of fragments produced, resulting in different banding patterns. This directly correlates with the specificity of the technique.

Examples of the use of this technique include identifying the hallucinogen *Salvia divinorum* using methods designed to discriminate it from its close relative *Salvia officinalis* (common sage). The potency of *S. divinorum* as an hallucinogen and its use as a 'legal high' have sparked discussion as to whether it should be banned. In order to do this effectively, a fast and simple identification technique would be required, and was developed using PCR-RFLP (Bertea et al., 2006).

The narcotic *Mitragyna speciosa*, Kratom, can also be identified using this method. This species has similar effects to opium and has been widely used medicinally. It has also been misused, which has led to its use being forbidden in many countries. Other morphologically similar *Mitragyna* species are used and sold in the place of *M. speciosa* creating a need for a reliable identification method (Sukrong et al., 2007).

Stemonae Radix is a widely used antitussive in TCM and is defined in the Chinese pharmacopoeia as either *Stemona sessilifolia*, *S. japonica* or *S. tuberosa.* The PCR-RFLP method was used to identify each of the three correct plant species, a closely related species *S. parvifolia* and a known adulterant *Asparagus cochinchinensis* (Fan et al., 2009).

Recently this technique has also been developed to separate Chinese Star anise, *Illicium verum*, from the toxic Japanese species *Illicium anisatum* (Techen et al., 2009).

Some prior knowledge of the sequence can be beneficial when utilising this technique. It is possible that a PCR amplicon will not contain any recognition sites for many endonucleases. Sequencing the target region in at least one species of interest can therefore save time and aid efficient experimental design.

In PCR-RFLP, the possible number of bands is capped due to the size of the PCR amplicon being used as the template. Direct RFLP begins with the restriction digest of genomic DNA; consequently the sequence variability is much higher as is the number of cleavage events. The banding patterns produced by this method are highly specific, but can be difficult to analyse due to the size and intensity of the bands produced. Genome restriction digests are often used as a precursor to other identification methods.

1.3.4 Amplified Fragment Length Polymorphism (AFLP)

AFLP is a robust technique which amplifies sequence differences which affect endonuclease recognition sites. An initial restriction digest of genomic DNA with two different endonucleases is followed by a ligation reaction in which double stranded DNA adapters are bound to the resultant cleaved recognition sites (Figure 1.1). A pre-selective PCR amplification stage is then carried out, the primers for which are an exact match to the adapter sequence with one other base added to the 3′ end, either A, G, C or T (Percifield et al., 2007). This causes a selection of approximately 1/16 of the available fragments for amplification, as only 1/4 of the sequences will match the base at the 3′ end of each primer. Another PCR then further selects from this population; the primers are fluorescently labelled and have a further three bases added to the 3′ end, so that again only the amplicons which match exactly will be amplified. Many primers with different additional bases and fluorophores may be used, separating fragments with different sequences directly flanking the adapters. Once a target specific marker has been found, it may then be used alone or in conjunction with many others. The end product of this technique is then a fast, reliable and simple PCR test, whether it is for one or many of the polymorphic bands. However, the intricacies of this technique can be time consuming and technically demanding.

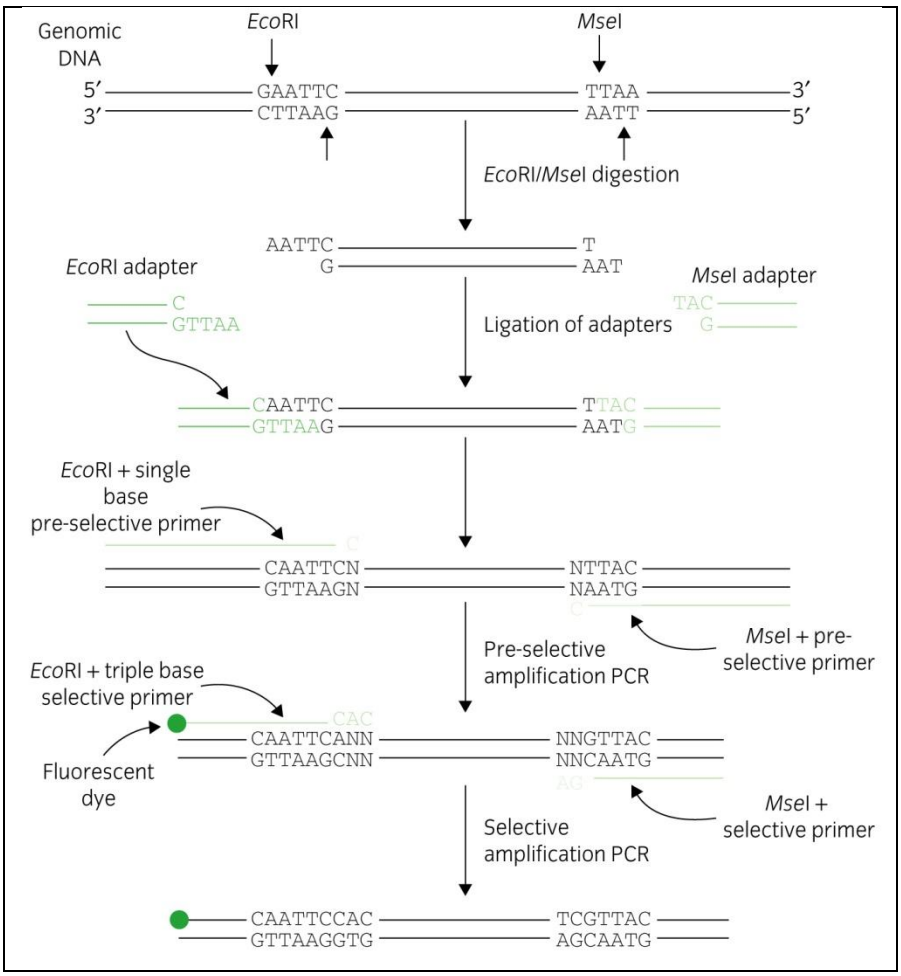


Figure 1.1 Diagram showing AFLP process.

The applications of this technique to medicinal plant authentication include separating *Hypericum perforatum* from other *Hypericum* species, and using the resultant polymorphic markers to construct phylogenetic trees (Percifield et al., 2007). The species *Actaea racemosa*, Black Cohosh, can also be identified using this technique, along with the possible adulterants *A. podocarpa*, *A. pachypoda* and *A. cordifolia*. The utility of the technique was also demonstrated with commercial samples which were shown to contain only *A. racemosa* (Zerega et al., 2002).

This technique has also been developed to identify each of three economically important *Echinacea* species: *E. purpurea* (the most valuable), *E. angustifolia* and *E. pallida*. All are used traditionally by Native Americans, and the growing plants can be readily discriminated from one another based on morphology. However, the most expensive species *E. purpurea* is often sold as dried crushed plant material, or in tablet form, making it indistinguishable from the other two species. The AFLP system allows material at any stage of processing to be identified (Russi et al., 2009).

1.3.5 Amplification-Refractory Mutation System (ARMS)

The ARMS technique depends on the analysis of sequences from different targets in specific regions. It is most useful in separating very similar individuals, such as closely related plant species. An initial PCR amplification is followed by sequencing of the amplicon. These data are then analysed for single bases which differ between the target sequences: Single Nucleotide Polymorphisms (SNPs). A series of primers are then designed to match the different versions of the sequences. The primers must match the single base difference at the 3′ end, this will ensure that the reaction will succeed or fail using just this one base difference as this base is essential for amplification. These are then used in a highly stringent PCR with the initial amplicon used as the template. Several primers are designed for different SNPs, so that with a selection of primers, the resulting gel pattern will reveal the identity of the target in question. The design and optimisation of these primers can be technically demanding and time consuming. The parameters necessary for a single base difference to prevent PCR product formation need to be very stringent and are essential to the success of the technique.

This method has also been applied to medicinal plants, allowing different levels of taxonomic identification. ARMS has been developed to discriminate a high yield Korean cultivar of *Panax ginseng* "Chunpoong", which was found to have a specific SNP when compared to other *Panax ginseng* cultivars (Wang et al., 2010).

Further uses of the technique include the ability to determine the geographical origin of specimens. Population specific SNPs were identified for *Dendrobium officinale*, enabling the classification of fifty market place samples (Ding et al., 2008), along with authenticating the samples as *Dendrobium officinale.*

The plant species *Alisma orientale* used traditionally in the treatment of urinary and kidney problems in TCM, can also be authenticated using this method in conjunction with PCR-RFLP marker patterns (Li et al., 2007). DNA sequences were compared with six other *Alisma* species to identify species specific SNPs which were then used for ARMS primer design.

1.3.6 Simple Sequence Repeats (SSRs)

SSRs are very short sequence motifs, 1-6 bases in length, which are repeated a number of times. They are the result of errors in the DNA replication process and occur throughout the genome, though predominantly in non-coding regions. The variability in length and number of SSRs enables their use for the purposes of identification.

When the sequence directly upstream and downstream of a particular SSR is known, primers may be designed to amplify these repeat regions. Samples are then characterised according to the length of the different repeat regions, via gel electrophoresis. As this technique requires knowledge of the target DNA sequence, this must be obtained prior to application.

To avoid the necessity for prior sequence knowledge, the SSRs may be analysed via another approach: Inter-Simple Sequence Repeats (ISSR). In this procedure, PCR primers are designed to anneal to the SSR regions in an outward orientation. In a similar fashion to RAPD, amplicons are formed by the chance annealing of these primers in an orientation to and at a distance from each other which allows the progression of the reaction. The result is an amplification of the regions in between adjacent SSRs, which again vary in length and number and may be analysed on that basis.

The applicability of this method without prior sequence knowledge increases convenience, but is balanced by the need for careful reaction optimisation, though it is slightly more reliable than RAPD which is dependent on similar optimisation procedures (See section 1.3.1). The ISSR technique has been used for the identification of 31 *Dendrobium* species, and to assess genetic diversity and relationships across the genus (Wang et al., 2009). The SSR method has also been developed for use in identifying different green tea cultivars (Ujihara et al., 2009).

1.3.7 Direct Sequencing

Genome sequencing efforts can provide absolute assurances of authentication and identification. However, sequencing an entire genome is time consuming and expensive, so the complete sequence is only known for a handful of plant species. Thus, the sequencing of an entire genome in order to identify an individual or a sample of an individual is currently extremely impractical.

1.3.8 Advantages of DNA based Techniques

One of the main advantages of molecular identification is the negligible amount of plant material required. In order to extract DNA, typically just 0.1g fresh sample is necessary. Once DNA extraction is complete, just a small amount of this is then needed for most assays, enabling the storage of the remainder of these samples. DNA samples can be much more readily stored than plant material (required for morphological references) due to their small size and inactivity when frozen. Electronic access to database information is also possible; with internet access and international data sharing efforts, a wealth of DNA sequence information is available with which to compare any sample. Molecular information does not

suffer from the variability due to plant age seen in morphological methods and is not dependent on singular chemical marker compounds (Crockett et al., 2004, Bertea et al., 2006, Techen et al., 2004, Wesselink and Kuiper, 2008, Sukrong et al., 2007). These features mean that DNA based methods can identify the species of medicinal plant material where chemical and morphological method cannot.

1.3.9 DNA Barcoding

The Consortium for the Barcode of Life (CBOL) is an international collaboration of natural history museums, herbaria, zoological collections, botanical gardens and university departments dedicated to the progression of DNA barcoding (www.barcoding.si.edu). Genetic barcoding presents the possibility of identifying all species by analysis of a specific region of the genome (Hebert et al., 2003, Kress et al., 2005). CBOL aims to develop a global standard for the identification of all known biological species (www.barcoding.si.edu).

The principle of DNA barcoding was first described by Hebert et al. (2003) and is based on the use of a short DNA sequence from a defined position in the genome as a unique identifier of a species (Hebert et al., 2003). This initially courted controversy as it was perceived to be an attempt to replace traditional taxonomic methods with molecular techniques, 'DNA taxonomy' (Ebach and Holdrege, 2005). This was refuted in several publications, (Schindel and Miller, 2005, Gregory, 2005, Hebert and Gregory, 2005, Cowan et al., 2006). The intention of DNA barcoding is to be used to identify individuals as belonging to a species, not for delineating species, and specimens used to create the barcode should always be vouchered by a traditional morphological taxonomist before DNA sequencing begins. DNA barcoding can be of great use to taxonomists, for instance acting as a sieve to filter large collections into groups to expedite inventories and analyses, and particularly to identify species when morphological features are no longer present (Schindel and Miller, 2005).

There are several advantageous features that are necessary and/or desirable in candidate barcode regions:

- The sequence must be variable enough between species to allow identification, but must also be sufficiently similar in each individual of a species to enable recognition.
- The area or 'loci' must be accessible in all species, so the region should be amplified using generic primers in all species. This requires that a sequence close to either side of the barcode is extremely similar in all species.

- They should not display any of the characteristics known to complicate sequencing, e.g. long repeats.
- The length of the sequence should be such that it may be sequenced in one reaction.
- Comparing the sequence to those from other species should be uncomplicated; this excludes sequences that differ from each other due to large or numerous deletions or insertions. These would then have to be accounted for and characterised individually before assessing differences between species.
- The sequence must be retrievable from degraded samples, such as those commonly collected by forensic professionals. These samples may be dried, aged, extremely small etc., so the barcode must be small enough to be less likely to be damaged, and abundant enough to be gained from a very small amount of sample. (http://www.kew.org/barcoding/rationale.html)

The DNA barcoding of animals and fungi has rapidly advanced using the "Folmer Region" of the mitochondrial cytochrome c oxidase gene, *cox1* (Hebert et al., 2003). This region fulfils the majority of the preferred characteristics explained above; it is 648 bases in length making it amplifiable and sequencable in one reaction. Highly conserved regions, allowing the use of generic primers and straightforward amplification, flank the sequence. Finally, the variability of the DNA sequence is of a frequency that allows species identification. Consequently, efforts to sequence this region in animals and fungi are well under way. The product of these efforts is the Barcode of Life Data Systems (BOLD) digital library (http://www.boldsystems.org/views/login.php). Users submit barcode sequence data from vouchered specimens, along with raw electropherogram traces, photographs of the sample and details of the collection process and location. This ensures a reference library of the highest quality. Later users may then submit their DNA sequences in order to assign their sample to a species as designated by the search results.

The *cox1* region is highly conserved in plants rendering it ineffective in barcoding initiatives (Kress et al., 2005, Chase et al., 2005). Research into potential barcode regions has continued separately for plants due to this. The mitochondrial genome as a whole was discounted as it is re-structured rapidly in plants (Kress et al., 2005).

In 2005, Chase et al. favoured the nuclear ribosomal Internal Transcribed Spacers (nrITS) used in conjunction with one or two plastid regions. The use of two regions, one from the nuclear

and one from the plastid genome, allows for the detection of genetic differences inherited from both parents, as the plastid genome is usually maternally inherited. The recommendation from Chase et al. was based on available sequence data in GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), and was identified as a short term solution which would require further work in the future (Chase et al., 2005). Kress et al. (2005) made an initial selection in the same year derived from a two tier selection process. A comparison of *Atropa belladonna* and *Nicotiana tabacum* yielded regions to be put forward for empirical testing with 99 species from 80 genera. This resulted in nrITS and the plastid spacer region *trnH-psbA* being recommended, due to the amount and variability of sequence data available (Kress et al., 2005). It was noted that plastid sequences are particularly desirable as barcodes due to the high copy number per cell, conferring ease of amplification in aged or damaged samples.

A study of 27 species of Cycads found that none of the putative barcode regions were satisfactory. Most effective was the nrITS, but difficulties, due to length variability (and therefore alignment problems) and allelic variation, made this region unacceptable (Sass et al., 2007). Chase et al. (2007) embarked upon a large scale review of available sequence data; the outcome of this was to recommend two possibilities, each of which contained three barcode regions all from the plastid genome:

Option 1: *rpoC1*, *rpoB* and *matK*
Option 2: *rpoC1*, *matK*, and *trnH - psbA*

The regions *rpoC1* and *rpoB* both encode subunits of the plastid-encoded plastid RNA polymerase, and *matK* encodes an RNA maturase. The use of multiple regions was suggested to balance deficiencies and desirable features in each; *rpoC1* and *rpoB* showed universality of amplification but low resolving power, *matK* demonstrated high sequence variation but available PCR primers required work to achieve universal amplification, and *trnH-psbA* sequences were problematic to align due to considerable length differences in the region between species (Chase et al., 2007).

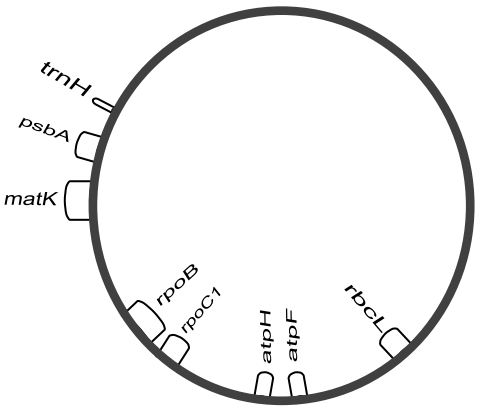


Figure 1.2 Chloroplast genome map showing candidate barcode regions

In 2007, a new experimental design using paired species in 48 genera was used to identify candidate barcode regions which were then tested for identification of species via GenBank (Kress and Erickson, 2007). The recommendation in this case was *trnH-psbA* along with a portion of the ribulose-bisphosphate carboxylase (RuBisCo) coding region in the plastid genome, *rbcL*. Previous difficulties with the alignment of *trnH-psbA* were discounted as this is mainly a problem for phylogenetic inference, which is not the aim of DNA barcoding. Newmaster et al. (2008) agreed with the use of *trnH-psbA*, but in conjunction with *matK* when analysing barcoding region sequences of 8 species of *Myristicaceae*, 40% of those known to science (Newmaster et al., 2008).

Lahaye et al. (2008) published analyses of >1600 samples from Costa Rica and the Kruger National Park in South Africa, both Biodiversity hotspots, mainly consisting of *Orchid* species. They recommended the *matK* region, as they did not encounter the amplification difficulties previously reported. *trnH-psbA* was also singled out as similarly applicable, with other plastid coding regions discounted due to high conservation of sequence. The likelihood of requiring a further region from the nuclear genome (due to uniparental plastid inheritance) was also highlighted.

In the same year, Fazekas et al. compared the discriminatory power of eight plastid barcode regions and the mitochondrial *cox1* region in 92 species from 32 genera (Fazekas et al., 2008). One of the plastid regions, the 23S rDNA coding region, and the mitochondrial *cox1* gene were eliminated from further consideration in the study due to the very low resolution found in the initial sequencing results. Of the remaining regions, three coding (*rbcL*, *rpoB* and *matK*) and two non-coding (*trnH-psbA* and *atpF-atpH*) were recommended as a short list. However, severe amplification difficulties were reported for *matK*, as were sequence inversion events for

trnH-psbA, and both of these regions contained homopolymer runs making sequencing laborious. The optimal number of barcoding regions for identification to the species level was demonstrated to be three or four, the authors pointing out that most of the cost of barcoding is due to sample collection and processing rather than PCR and DNA sequencing, so increasing the number of barcode regions would not dramatically affect the cost.

In 2009 CBOL published a report from the pooled data of 445 angiosperms, 38 gymnosperms and 67 cryptogams. The *trnH-psbA* region showed 93% amplification and good discrimination, though low quality sequence data and lengths varying from 300 to 1000 base pairs in different species caused it to be removed from consideration. A two region system was advocated: *rbcL* as it is the most characterised region and provides good universality and sequence data (although it is not highly variable) and *matK* due to the high resolving power which balances its limited amplification in gymnosperms and cryptograms (Hollingsworth et al., 2009, CBOL et al., 2009).

Two proposals were submitted to CBOL to name the plant DNA barcodes; one based on the three locus system described by Chase et al, and the other based on CBOL's own two locus system. The Executive Committee found the three locus solution scientifically valid, but cautious, and viewed it as unlikely to show great advantage over the two locus proposal. In addition to this, the delays and cost implications of naming three regions as plant barcodes were expected to unduly hinder the advancement of the initiative. The two locus approach was therefore accepted by the Barcoding of Life Initiative, and sequences of these two regions can now be submitted to GenBank with the reserved keyword "Barcode". CBOL will review this position in 18 months to assess expected advances in the universality of primers for the *matK* region, and in the alignment of *trnH-psbA* region sequences (Executive Committee, 2009).

Recently, however, the ITS2 sequence of the nrITS region has been proposed as a barcode for Chinese Medicinal Plants (Chen et al., 2010). A sequence analysis of >6600 plants from 4800 species listed in the Chinese Pharmacopoiea showed that this region was more applicable than all others previously described as putative barcodes. Thus, identification of the barcode regions for plant species, and whether these will be the most appropriate for authentication purposes, remains a contentious issue.

1.4 *Hypericum perforatum* L., St John's Wort

St. John's Wort (SJW), *Hypericum perforatum*, is an herbaceous perennial with characteristic translucent dots on the leaves when held to the light. It is native across Europe, Asia and North Africa and has been introduced to many other temperate regions such as America and Australia where it has become abundant (Southwell and Bourke, 2001). Traditionally it has been used for centuries in the treatment of many conditions, ranging from burns and rheumatic joints to asthma (Hunt et al., 2001).

Today *H. perforatum* is predominantly used to treat mild to moderate depression, and is prescribed for this purpose within the European Union (McGarry et al., 2007). It is currently one of the most widely used medicinal plants in Europe, contributing $40 million to the German Health Insurance reimbursement figure in 2003 (De Smet, 2005) and has been increasing in popularity dramatically in the United States of America, with annual sales rising from US$20 million to US$200 million between 1995 and 1997 (Gaster and Holroyd, 2000).

1.4.1 Efficacy

SJW is primarily used in the treatment of mild to moderate depression, and has been the subject of many trials to determine its efficacy compared to both placebos and pharmaceutical anti-depressants. The Cochrane review, a systematic review and meta-analysis of 29 trials; 18 against placebo and 17 against standard antidepressants, was conducted in 2008 (Linde, 2008). All studies selected for inclusion in the review were double-blind, randomised controlled trials of the highest standard. Each study was assessed as to whether SJW was more effective than a placebo, as effective as a typical antidepressant and whether it caused fewer adverse effects.

SJW was found to be more effective than a placebo; those taking SJW were 28% more likely to respond to treatment. In comparison to pharmaceutical antidepressants, SJW showed a similar level of efficacy, and adverse effects, as measured by drop-out rate, were found to be lower for SJW (Linde, 2008).

A finding which has been more difficult to explain is that in all cases, results from German speaking countries were more favourable towards SJW than others. Although use of phytopharmaceuticals in Germany is widespread, this should not affect double-blind randomised trials. The authors interpreted the cause of this as either a difference in the diagnosis of patients in these countries, particularly as this is carried out by private practitioners in Germany as opposed to hospital or academic staff, or the inclusion of smaller

trials which may have overoptimistic results (Linde, 2008). The likelihood of significant bias in the results from the review remains low despite this.

Since this review, SJW has also been shown to prevent relapse after an acute depressive event, while tolerability and adverse events were near to or at the placebo level. The relapse rate was shown to be reduced over a period of 6 months after the initial event (Kasper et al., 2008).

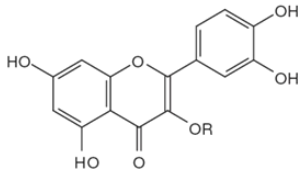
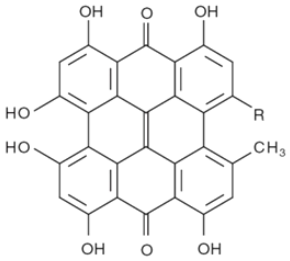
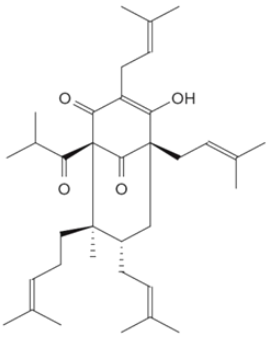
A subsequent meta-analysis of studies between 1966 and 2008 compared the efficacy and tolerability of SJW to the group of standard anti-depressants termed selective serotonin reuptake inhibitors (SSRIs) in the treatment of major depressive disorder. Placebo controlled trials were selected, a total of 13 studies. The results indicate that SJW was as effective as SSRIs, and also resulted in fewer adverse event related withdrawals from the studies (Rahimi et al., 2009). This is clearly preferable in the treatment of depressive disorders.

Though these reviews report positive findings about the efficacy and safety of SJW, most also warn of the drug-interactions known to occur (see section 1.4.3) and also the quality of SJW products on sale. The composition of products must be considered before studies are conducted, as they can vary dramatically.

1.4.2 Active Compounds

Of the many constituents of SJW, hypericin was the first to attract attention as the putative active compound causing the anti-depressive effect of SJW, the structure of which is shown in Table 1.1. In a forced swimming test (FST), a common model system for assessing depression in rodents by measuring immobility time on introduction to water, both hypericin and pseudohypericin were found to be positively effective. Interestingly, both were found to be significantly more effective in the presence of another fraction containing procyanidins which solubilised hypericin. Investigations using a dopamine antagonist, sulpiride, showed antagonism of the effects of hypericin and pseudohypericin, suggesting that both act via the dopaminergic system (Butterweck et al., 1998).

Table 1.1 Chemical Structures of the main compounds involved in the antidepressant activity of SJW, adapted from (Butterweck, 2003).

Group of bioactive compound	Structure	Constituent
Flavonoids		1. R = H Quercetin 2. R = α-L-rhamnosyl Quercitrine 3. R = β-D-glucosyl Isoquercitrin 4. R = β-D-galactosyl Hyperoside 5. R = β-D-rutinosyl Rutin 6. R = β-D-glucuronide Miquelianin
Naphthodianthrones		1. R = CH_3 Hypericin 2. R = CH_2OH Pseudohypericin
Phloroglucinols		Hyperforin

Hyperforin, a phloroglucinol shown in Table 1.1, has also generated attention as a possible antidepressant compound. It had previously been discounted due to its instability, inferring that it cannot be present at sufficient concentrations to cause the antidepressant effect of SJW preparations as it is labile. However, SJW is often available in tinctures (alcoholic extracts) in which hyperforin would be more stable. Preparations of SJW extracts with extremely high hyperforin concentrations, and without hypericins and other compounds, were found to inhibit uptake of serotonin in peritoneal cells. These extracts also showed antidepressant activity in behavioural despair tests, another term for the FST (Chatterjee et al., 1998).

Flavonoids from SJW have also been shown to be biologically active. An HPLC fraction containing hyperoside, isoquercitrin, miquelianin and quercitrine, and another with hyperoside

and astilbin among many unknown compounds, were tested in the FST. All of these compounds, excluding quercetin, quercitrine and astilbin, were found to be active in reducing immobility times (Butterweck et al., 2000).

This is compounded by the finding that SJW extracts without hypericin or hyperforin still exert an antidepressant effect in the FST and also the tail suspension test (TST), an antidepressant model in mice. This supported the idea that flavonoids are a factor in the antidepressant effect, as the removal of hypericin and hyperforin resulted in an extract with higher than usual concentrations of flavonoids (Butterweck et al., 2003).

These three groups; napthodianthrones, phloroglucinols and flavonoids, are generally accepted as having biological function with regard to antidepressant effect. However, as none of these groups alone can account for the activity of SJW, it is largely agreed that a whole preparation containing all compounds found to occur in SJW is most effective, and must therefore be considered the active substance (Kober et al., 2008, Butterweck and Derendorf, 2008).

1.4.3 Drug Interactions

SJW is known to interact with pharmaceuticals, this is thought to be due to induction of the transcription of cytochrome P450 (CYP) enzymes by hyperforin (Saxena et al., 2008, Moore et al., 2000). CYP proteins are a superfamily of enzymes found in all living things; they are involved in many functions including the metabolism of pharmaceuticals in humans. Altering the activity of these enzymes can result in a change in the rate at which drugs are degraded and cleared from the body. SJW preparations have been shown to increase this activity, resulting in an increased rate of clearance and consequently reduced effect of the pharmaceutical. This means a higher dose is needed to cause the desired effect.

A number of pharmaceuticals are known to be affected by SJW in this way: cyclosporine, an immunosuppressant administered to transplant patients (Saxena et al., 2008), omeprazole, a treatment for gastric ulcers (Saxena et al., 2008) and imatinib and irinotecan, anti-cancer medications (Kober et al., 2008).

This could also explain loss of function episodes observed in other pharmaceuticals when SJW is a factor, such as warfarin, a widely used anticoagulant (Saxena et al., 2008), and ethinylestradiol, an oral contraceptive for women (Saxena et al., 2008).

It has been suggested that hyperforin levels in SJW products should be capped, for instance at 1 %, to prevent these CYP450 herb-drug interactions (Butterweck and Derendorf, 2008). This is supported by findings that the observed reduction in oral contraceptive efficacy does not occur with low hyperforin content SJW preparations (Will-Shahab et al., 2009). Currently, dried extracts are required to contain no more than 6% hyperforin (European Pharmacopoeia, 2008).

The interactions of SJW are of particular concern, as consumers appear to be largely unaware of them. Indeed, given that 58% of UK consumers believe that 'herbal medicines are safe because they are natural' it is not remarkable that 22% of them feel that telling their GP about this use is not necessary (Medicines and Healthcare products Regulatory Agency, 2009). This may be due in part to the lack of consistent warnings on SJW product labels (Clauson et al., 2008).

Overall SJW has been shown to be an effective anti-depressant with few side effects. In comparison to many pharmaceutical anti-depressants the side effects which are reported are low risk and low occurrence. Coupled with the general increase in the usage of medicinal plants, this indicates that SJW based products will become more popular.

1.4.4 Morphological and Chemical Identification

1.4.4.1 Morphological Identification

The European Pharmacopoeial Monographs (European Pharmacopoeia, 2008) contain the information necessary to identify and manufacture Medicinal and Pharmaceutical Substances. In the St John's Wort section, a detailed description is given of the plant's features which include the following:

- Leaves are 15-30 mm in length, opposed, and without stalks or stipules (Figure 1.3).
- Small black dots are seen on the outer edges of the leaves, and many small translucent glands cover the surface of the leaf (Figure 1.4).
- The flowers have 5 yellow petals with black glands and three stamina blades each with many orange yellow stamen (Figure 1.5).

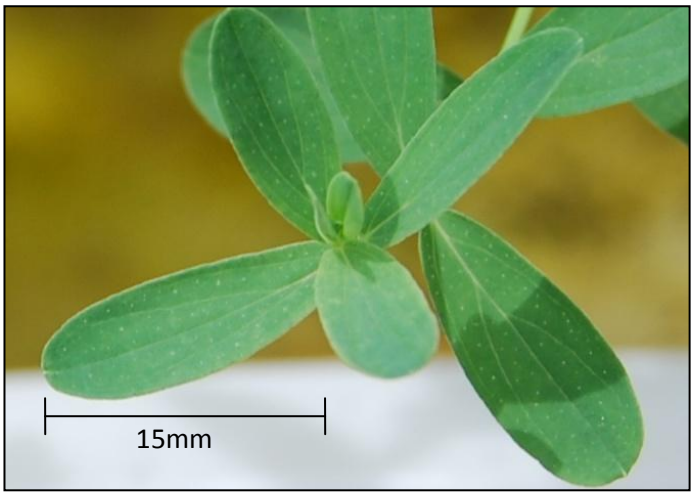


Figure 1.3 *H. perforatum* plant after 10 months growth, leaves showing morphological features of translucent dots spaced over the entire surface. Picture taken in July 2008.

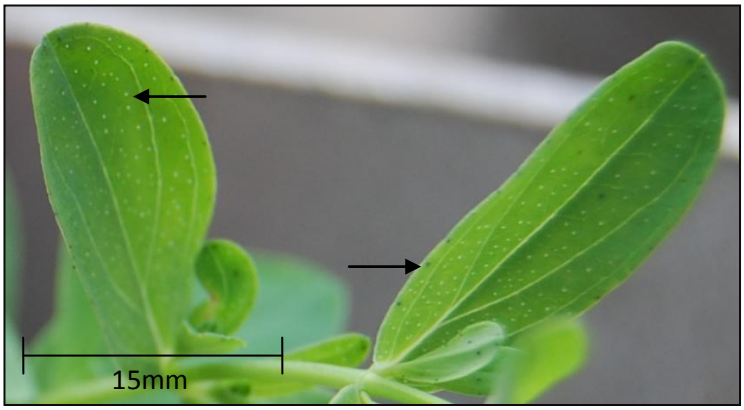


Figure 1.4 *H. perforatum* plant after 10 months growth, leaves showing translucent dots, and black dots which are glands containing hypericin (indicated by arrows). Picture taken in July 2008

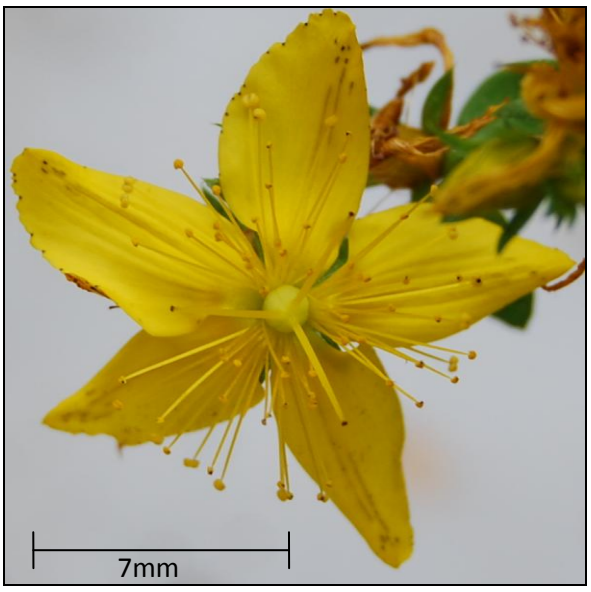


Figure 1.5 Star shaped *H. perforatum* flower on plant after 10 months of growth. Small black regions are visible on the petals, containing hypericin. Taken in July 2008.

The microscopic diagnostic features of SJW are listed by the European Pharmacopoeia as the following, when examined using chloral hydrate solution;

- Secretory glands containing essential oils, transparent (Figure 1.6).
- Secretory glands containing hypericin, dark red coloured (Figure 1.7).
- Cluster-crystals of calcium oxalate (Figure 1.8).
- Spiral form vessels (Figure 1.9).
- Epidermis with necklace shaped walls (Figure 1.10).

Figure 1.6 *H. perforatum* gland containing essential oil, pink background colour is due to the hypericin containing glands having burst on heating. Magnification x10.

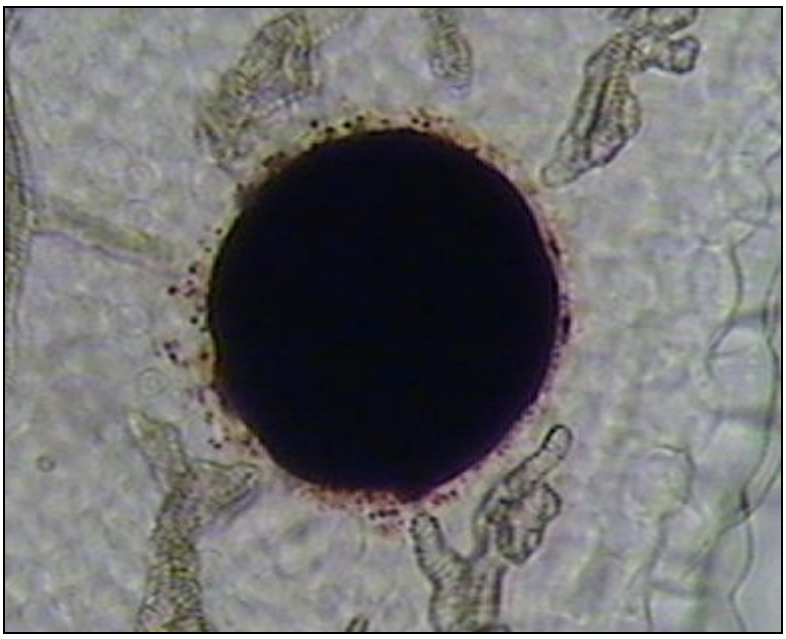


Figure 1.7 Gland containing hypericin in *H. perforatum*, the characteristic deep red colour appearing almost black. Magnification x10.

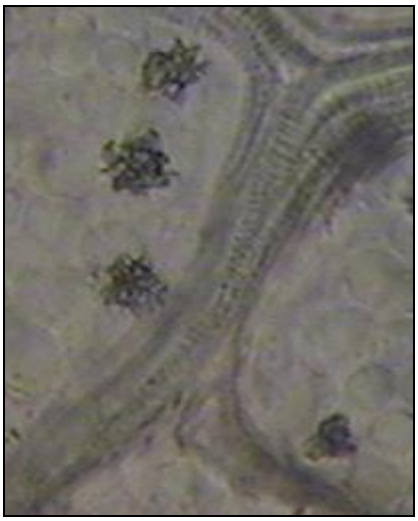


Figure 1.8 Calcium oxalate crystals in *H. perforatum.* Crystals were visible in many different areas all over the leaf. Magnification x10.

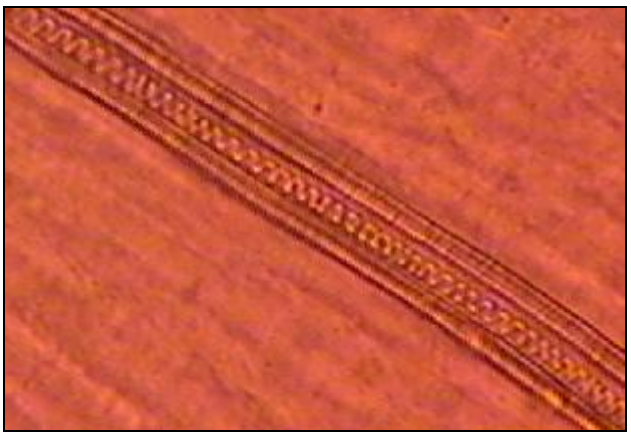


Figure 1.9 *H. perforatum* spiral form vessels clearly visible along the interior of the vessel. Magnification x10

Figure 1.10 Beaded feature of *H. perforatum* cell walls, also described as necklace forms. Magnification x10.

These identifying features are not, however, specific to *Hypericum perforatum*. They are found in many other *Hypericum* species which may be mistaken for *H. perforatum* or used as an adulterant.

1.4.4.2 Chemical Identification

The European Pharmacopoiea sets out two monographs for SJW (*H. perforatum*): one detailing requirements for the plant material and another for extracts thereof.

The chemical content of *H. perforatum* plant material is defined as a minimum of 0.08% of total hypericins, identified using TLC. The chromatogram results are described, with areas and colours of fluorescence indicative of rutin, hyperoside, pseudohypericin and hypericin. The

measure of percentage content is then discovered via the absorbance of an extract. Dried extracts must adhere to different requirements, based on a definition detailing hypericins, flavonoids and hyperforin. Compounds are identified via TLC, and assayed via liquid chromatography with separate procedures for hypericin and another for flavonoids and hyperforin.

2 DNA Based Identification of Medicinal Plant Material by PCR Primer Design

2.1 Introduction

2.1.1 The Array of Available Techniques

The number and variety of available techniques for DNA based identification of medicinal plants has led to the area becoming complex and confusing, with different research groups favouring particular techniques and applying them to their plant species of interest. Indeed, many of the aforementioned techniques may have been developed for one medicinal plant species by different groups on separate occasions, possibly even for different purposes. For instance, the important TCM plant genus *Dendrobium* has been analysed for different species via RAPD (Zha et al., 2009), ARMS (Ding et al., 2008) and DNA barcode sequence (Asahina et al., 2010).

In order for the potential of DNA based identification techniques to be realised, they must first become standardised between plant species and working groups (Cowan et al., 2006). DNA barcoding aims toward this 'gold standard', but delays in the choice of the barcode region have so far hindered this. Now that the decision is made (subject to review in 18 months) efforts to sequence *rbcL* and *matK* in all available plant species will no doubt follow. However, CBOL also advocates the sequencing of several other regions, previously candidate barcodes, to 'back-up' *matK*, and further work on procedures to improve amplification and sequencing of these regions (Executive Committee, 2009). The result of this will be a wealth of DNA sequence knowledge for between two and seven regions in plant species.

In order to make use of this resource to identify unknown plant materials, samples must first be subjected to DNA extraction, PCR amplification, amplicon clean-up and sequencing. This process, excluding DNA extraction, is required for each region chosen. Though becoming more manageable with technological advances, DNA sequencing is currently still a process which requires expertise and a specialised working environment and equipment. A simpler technique, more accessible to industry and regulators could increase the use of DNA based identification techniques.

2.1.2 Primer Design to Microcodes

The data produced by DNA barcoding initiatives, instead of being stored for reference, could be used to aid yet quicker and easier methods of identification. These data provide a unique platform to revolutionise the perception of the 'pro's and con's' of different DNA based identification techniques, and an opportunity to design new ones.

The theory behind the design of this technique is based on the observation of Summerbell et al. (2005) that within a barcode region there are "microcodes", regions of less than 25bp which contain adequate sequence variation to distinguish one species from related species (Summerbell et al., 2005). This idea was developed based on the high throughput screening method of micro-arrays, a system in which immobilised oligonucleotides are used as probes to bind complimentary DNA in samples. This system requires that the oligonucleotides bind only the intended target DNA, and bioinformatic analysis of barcode sequence data is used to identify these microcode regions which show the greatest probability of species specificity. Within a barcode region there could be several microcode regions, each represented within a microarray enabling certainty as to sample identity.

The ability of short DNA sequences, down to 150bp, to identify species was shown by Meusnier et al. (2008). Focusing on DNA sequencing for assay design, analysis of barcode data for many different animal species indicated that the relationship between sequence length and species resolution reached a plateau between 100 and 200bp (Meusnier et al., 2008). This indicates that much smaller regions within the entire barcode could be used to identify species, the microcode regions.

The suggestion in this research is that the design of PCR primers to these microcode regions will allow the identification of a target plant species, and that this will provide a model of how barcode information could be used in the future to design sequence specific identification probes. This model would then have the potential to be reproduced and used as a model for the design of future investigations, in order to quickly and easily identify plants to the species level. The end-product of this would be simple PCR assays providing a fast positive or negative result.

2.1.3 Nuclear Ribosomal Internal Transcribed Spacer Regions (nrITS)

The nuclear ribosomal Internal Transcribed Spacer (ITS1 and 2) sequences are regions of genomic DNA flanked on the 5' and 3' ends by the coding sequence for 18S and 5.8S rRNA and

5.8S and 28S rRNA respectively (Figure 2.1). The rRNA coding regions are highly conserved throughout plant species; this allows the entire nrITS region to be amplified in a diverse range of plants using the generic primers ITS1 and ITS4 (White et al., 1990). Throughout this thesis the nrITS referred to is the entire region between these two primer annealing positions.

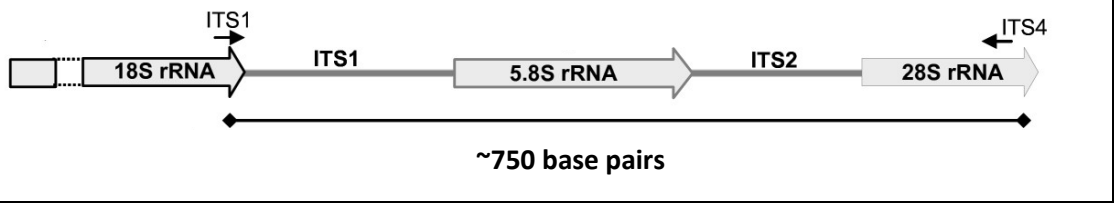


Figure 2.1 Representation of the nuclear ribosomal coding region with the Internal Transcribed Spacers shown, ITS1 and ITS2. The annealing positions of the amplification primers used are shown, ITS1 and ITS4, the region in between these primers is referred to as the nrITS throughout this thesis.

For the initial investigation of the theory of primer design to microcodes, the genomic nrITS regions were chosen. Initial studies into DNA barcoding found the nrITS useful (Kress et al., 2005, Chase et al., 2005). However, significant length differences between species, potentially causing alignment problems (Sass et al., 2007) along with complicated molecular evolution patterns (Chase et al., 2007) led to the decision against using the nrITS as a barcode region. These features affect the utility of the nrITS as a universal barcode, but are present only in some land plants (Chase et al., 2007), and recently these regions, or sections of them, have come somewhat back in to favour (Chen et al., 2010).

The nrITS sequences within *Hypericum* species have an evolutionary rate which results in inter-species variation and intra-species conservation (Crockett et al., 2004), enabling the distinction between species based on these regions. DNA sequence data for these regions of more than 50 *Hypericum* species was made available by Crockett et al. (2004), modelling the platform for PCR primer design which in future will be provided by DNA barcoding.

2.2 Materials and Methods

2.2.1 DNA sequence analysis

All available published *Hypericum* nrITS DNA sequences were obtained from the GenBank database (July 2007), with Accession Numbers and species as listed in Appendix section 8.1.1. Multiple alignments were carried out using the ClustalW program (Chenna 2003) hosted on the European Bioinformatics Institute (EBI) website (www.ebi.ac.uk), using the default settings. The Fingerprint program (Lou and Golding 2007) available at

http://evol.mcmaster.ca/fingerprint/ was used to display and manually optimise the alignment, and calculate nucleotide variation within the nrITS regions.

2.2.2 Sample Materials

Vouchered DNA samples from the Royal Botanic Gardens, Kew, DNA Bank (http://data.kew.org/dnabank/homepage.html) were used as confirmed species specimens. Details of the species and Kew Voucher Numbers are as follows;

Table 2.1 Voucher numbers and species for samples from the Kew DNA Databank

Voucher number	*Hypericum* species	Authority
13854	*H. androsaemum*	L.
13866	*H. kouytchense*	H.Lév.
13876	*H. perforatum*	L.
13896	*H. maculatum*	Crantz
13908	*H. patulum*	Thunb.
13921	*H. perforatum*	L.
13923	*H. athoum*	Boiss. & Orph.
13929	*H. calycinum*	L.
13932	*H. perforatum*	L.
13938	*H. delphicum*	Boiss. & Heldr.
13993	*H. ascyron*	L.

Seven further samples were collected, four from seedling material and three from garden grown varieties. Fresh plant material was cultivated from seeds supplied by Chiltern Seeds Ltd. (Bortree Stile, Ulverston, Cumbria, U.K. LA12 7PB) for the following species; *H. perforatum* L. (Ref. 701C), *H. kouytchense* H.Lév. (Ref. 700J), *H. androsaemum* L. (Ref. 698E) and *H. ascyron* L. (Ref. 698N). Samples were taken from four week old seedlings of each species and subjected to DNA extraction.

Leaf samples were collected from garden varieties of *H. calycinum* and the Hidcote cultivar, plus one further Hidcote sample from a commercial supplier. This cultivar is often sold as 'St. John's Wort' in garden centres, though it is not *H. perforatum* but actually a hybrid of *H. calycinum* and *H. cyathiflorum*

(http://www.rhs.org.uk/Databases/HortDatabase.asp?ID=93054). These were all subjected to the DNA extraction process.

The commercial materials were capsules filled with dried, ground plant material from three companies, labelled as containing the following;

Company A – 334mg St. John's Wort Extract per capsule.

Company B – 300mg St. John's Wort Pure powdered herb per capsule.

Company C – 333mg St. John's Wort (*H. perforatum*) Standardised Herb Extract and 114mg of other plant extracts and concentrates.

2.2.3 DNA Extraction

DNA extraction was carried out utilising the Qiagen DNeasy® Plant Mini Kit and TissueLyser (Qiagen Inc., CA). The Mini Protocol was used, pages 24-27 of the DNeasy® Plant Handbook.

Samples were either 0.1g fresh shredded plant material, or 0.02g dried material from capsules. These were placed into 2mL safe-lock microcentrifuge tube with 400µL Buffer AP1 and 4µl RNase A and a 3mm tungsten carbide bead. The tubes were placed evenly in the TissueLyser Adapter Set, and subjected to two disruption steps of 1min at 30 Hz, the tubes were reversed in position between disruption steps to achieve equal homogenisation.

The Manufacturer's instructions were then followed from step 8, recommended centrifugation steps were followed and two elution steps of 100µL were performed, providing a total elution volume of 200µL.

2.2.4 PCR Protocols

The nrITS1 region was amplified using primers as follows:

ITS1 (5′-TCCGTAGGTGAACCTGCGG-3′)

ITS4 (5′-TCCTCCGCTTATTGATATGC-3′) (White et al., 1990)

The previous candidate barcode region *rpoC* was amplified using the primers made publicly available on the Royal Botanical Garden, Kew, website (www.kew.org/barcoding):

rpoC 2 (5′-GGCAAAGAGGGAAGATTTCG-3′)

rpoC 4 (5′-CCATAAGCATATCTTGAGTTGG-3′)

PCR primers designed were as follows, forward:

FO2 (5′-CATAAGAAGTGTAAGGCTCCCGG-3′)

FIn (5′-GACAACACGGTCGGGGGCCT-3′)

Reverse:

HRI-S (5′-AGAGTCGTTATTGTTATGAACAGAAGGAG-3′)

RO (5′-TTCTGCAATTCACACCAAGTATCG-3′)

HR373 (5′-CCATCCTATTCCCGATTGTCTCTT-3′)

PCR reactions consisted of Green GoTaq® Flexi Buffer (Promega, Madison, WI, USA) (1x), $MgCl_2$ (2.5mM), GoTaq® DNA Polymerase (Promega) (1.25 Units), relevant primers (0.1µM each), dNTPs (0.1µM each), and template DNA (0.7-1µg) made up to a final volume 50µL with nuclease-free water in 0.2mL polypropylene tubes. The Applied Biosystems GeneAmp PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA) was used with the programme: 7min at 95^{o}C initial denaturation step, 30 cycles consisting of 1min at 95^{o}C, 30s at 60^{o}C and 1min at 72^{o}C, final extension period of 7min at 72^{o}C. This programme was suitable for all primer combinations used. Cycle number was increased to 40 in detection level assays.

Reactions without template DNA were utilised as controls. PCR products were run on 3% (w/v) agarose, 0.5 X TBE gels with 2µL SYBRsafe™ (Invitrogen, Carlsbad, CA, USA) DNA stain at 90V for ~30min and analysed in a BioRad Illuminator with ChemiDocXRS Camera and Quantity One software.

2.3 Results and Discussion

2.3.1 Primer Design

A multiple alignment of all the available published nrITS sequences for *Hypericum* species in the GenBank database (July 2007) was carried out. The outputs from multiple alignment software conventionally highlight regions of identity, but do not show the extent of variability within regions which are not identical. Fingerprint Software (Lou and Golding, 2007) is a tool for the calculation and display of variability in a multiple alignment. Figure 2.2 shows two such measures of sequence variation, the number of variants and nucleotide diversity, calculated from a multiple alignment of 91 *Hypericum* nrITS sequences. The number of variants analysis counts the number of different nucleotides found at each base position, the maximum being four. Whereas nucleotide diversity is a measure of how many of the input sequences differ at any base position, with a maximum value dependant on the number of sequences analysed.

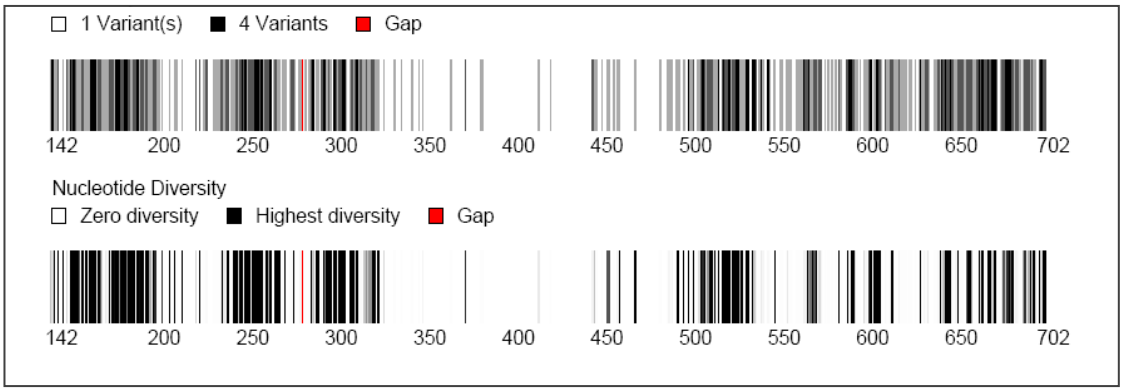


Figure 2.2 Fingerprint software output showing analysis of the 5.8S, ITS1 and 18S regions of 91 *Hypericum* species (Lou and Golding, 2007).

This highlighted areas of maximum divergence (potential microcodes) within the *Hypericum* nrITS sequences, indicated by the black regions in Figure 2.2. PCR primers were then designed to anneal to these microcode regions within the *H. perforatum* sequence, giving the highest probability of a species-specific PCR product. In total, five primers were designed, four intended to be specific (2 forward: Fin and FO2; 2 reverse: HRI-S and HR373) and one non-specific reverse primer (RO), to different areas within the nrITS region (Figure 2.3). These sequences were then checked against the nrITS sequences of the *Hypericum* species to be tested (Figure 2.3).

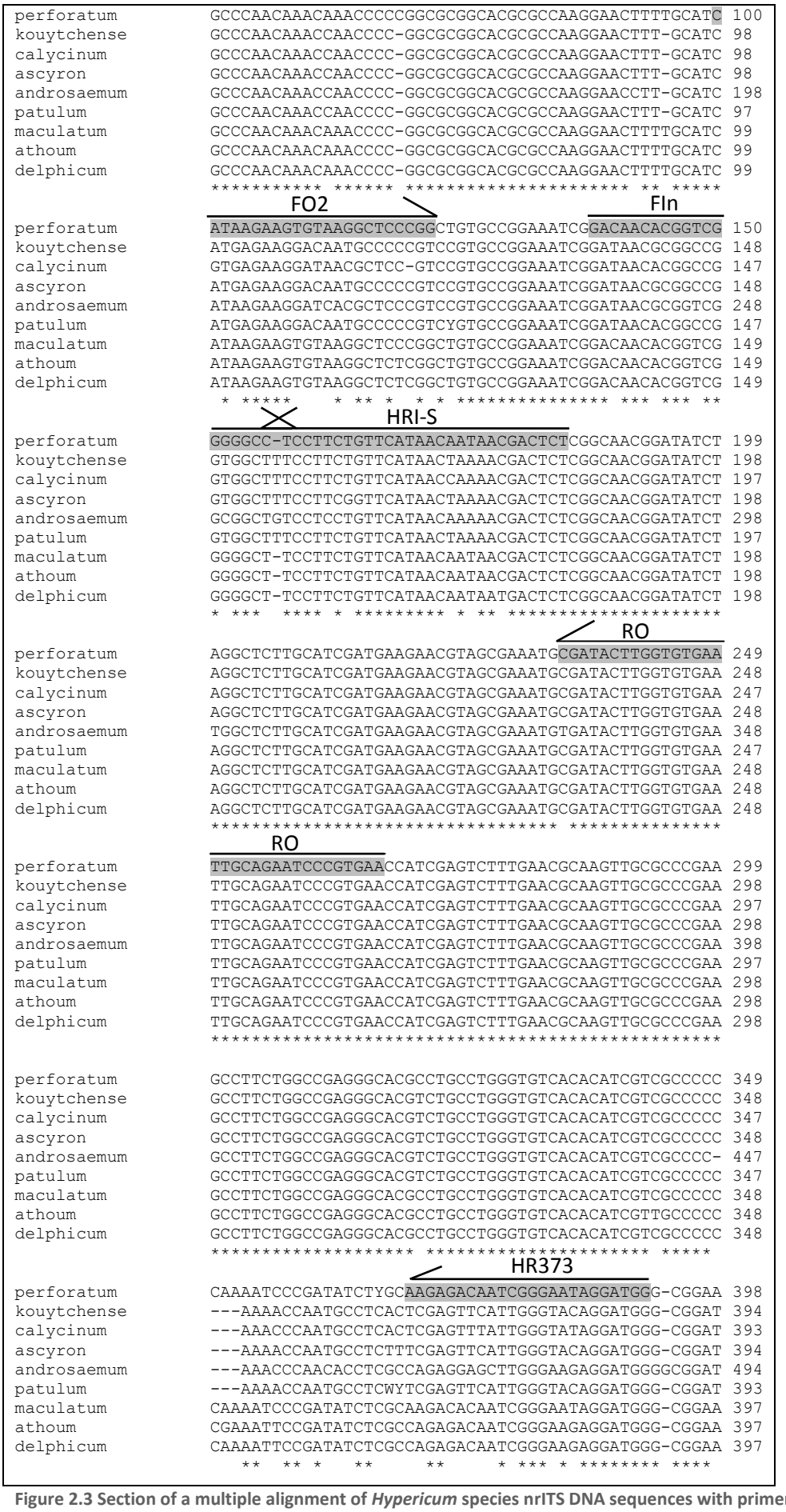


Figure 2.3 Section of a multiple alignment of *Hypericum* species nrITS DNA sequences with primer sequences and annealing positions indicated.

The four proposed primer combinations were all shown to be specific to *H. perforatum* when tested against four other *Hypericum* species, the results from two pairs being shown in Figure 2.4. Each primer combination has different qualities based on their length, sequence, GC content, and the amplicon length and sequence. The amplicon lengths produced using the reverse primer HR373 are slightly longer than preferred (255bp with FIn and 293bp with FO2). FIn contains a run of five G's flanked by three C's contributing to a 70% GC content which is not ideal as it increases the annealing temperature required and the probability of non-specific binding. Based on this, FO2 and HRI-S were selected as the most suitable primers to form a pair. The annealing temperatures of these primers are identical and FO2 has a GC clamp at the 3′ end, a run of three C's and two G's, which helps to promote specific binding and therefore amplification. This is in contrast to the long GC run in FIn, which is not at the 3′ end. The amplicon from this primer pairing is 85bp, which is at the lower end of the recommended amplicon length range for PCR.

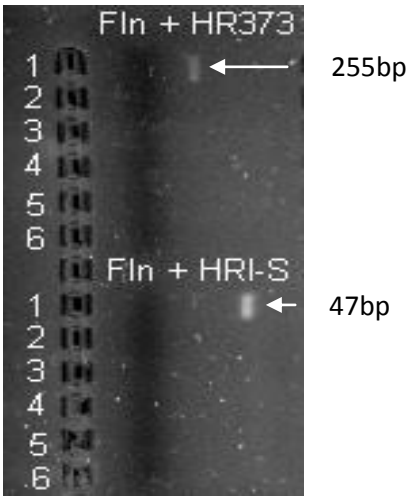


Figure 2.4 Image of a gel showing the *H. perforatum* specific products of two primer pairings as indicated.

Numbered lanes: 1 *H. perforatum*, 2 *H. kouytchense*, 3 *H. ascyron*, 4 *H. androsaemum*, 5 *H. calycinum*, 6 Negative Control.

A short amplicon can be beneficial when dealing with aged or highly processed plant material, as DNA degradation can occur. This can be caused by many factors including acid or alkali conditions, and the presence of DNA degrading enzymes within the plant material, and results in a reduction of the quality of all the DNA present. These conditions may occur as part of the manufacturing process of medicinal plant products, preventing the useful application of DNA-based identification methods. Another specific form of DNA degradation can be caused by the mechanical breakdown of the plant material, DNA shearing. In this process, the DNA molecule is broken at random intervals and, should breakage occur between the primer annealing sites,

amplification will not occur. This can be guarded against in the design of assays as the length of the amplicon increases the likelihood of this occurrence; so shorter amplicons are suited to these materials. As many medicinal plant products are highly processed, and often dried in non-ideal conditions for DNA preservation, the short amplicon produced by the FO2/HRI-S pair of primers was regarded as a valuable property.

2.3.2 Empirical Testing

The highly conserved 18S and 28S rRNA coding regions flanking the nrITS region allow the entire sequence to be amplified in a diverse range of plants using the generic primers ITS1 and ITS4. This enables a two tiered experimental design, the ITS1 and ITS4 primers verifying the presence of the nrITS regions within the genomic DNA extraction, and the FO2 and HRI-S pairing indicating the specific presence of the *H. perforatum* nrITS region.

Using the eleven Kew vouchered DNA samples, the primer pairing of FO2 and HRI-S gave positive result for each of the three *H. perforatum* samples producing the expected amplicon, showing consistency of amplification in individuals within a species (Figure 2.5). One other sample gave a product with the pairing of FO2 and HRI-S, *H. delphicum*. A pairwise comparison of the sequence at the primer annealing positions with *H. perforatum* shows a similarity of over 90% (Table 2.2), explaining the amplification. However, this species is not widespread (Crockett et al., 2007), and is unlikely to be found as a substitution or adulterant of *H. perforatum* on sale commercially. Two further samples have a comparable sequence similarity to *H. perforatum* at the primer annealing sites, *H. athoum* and *H. maculatum* (Table 2.2) though neither gave a product. The *H. athoum* sample gave no product with either the ITS1 and ITS4 combination or the FO2 and HRI-S combination. This indicates that amplifiable nuclear DNA was not present in the sample either due to a PCR inhibiting agent or low quality DNA. Amplifiable nuclear DNA was present in the *H. maculatum* sample. The negative result with FO2 and HRI-S may be due to sequence differences not shown in the published material. There are two published sequences for the nrITS regions in *H. maculatum*, AY573007 and AY555842. These differ in sequence in the primer annealing position of FO2 (Figure 2.6), with 95.7% and 100% sequence similarity to the *H. perforatum* sequence respectively (Table 2.2). This highlights the variability of published sequence data and, as no product is found experimentally, also calls into question the accuracy of sequence data in public databases.

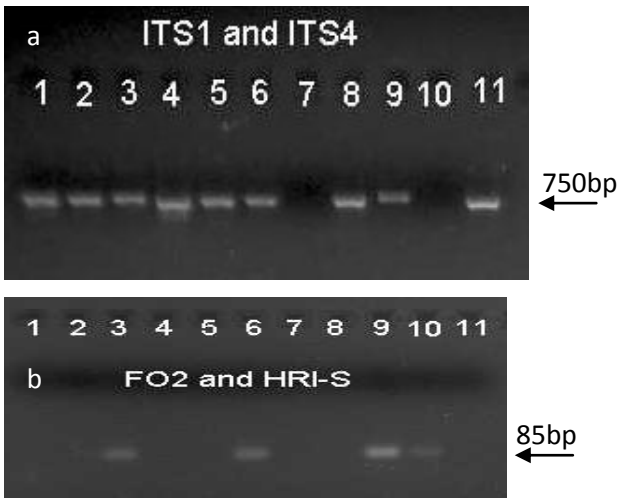


Figure 2.5 Image of gel showing results for vouchered DNA samples with two primer pairs.

a, PCR products from primers ITS1 and ITS4 with vouchered specimen *Hypericum* DNA samples from Kew DNA Databank. Products are clearly visible in all lanes other than 7 *H. athoum* and 10 *H. delphicum*. PCR products from primers ITS1 and ITS4 with vouchered specimen *Hypericum* DNA samples from Kew DNA Databank. Fig. 2.5 b, PCR products from primers FO2 and HRI-S with Kew DNA samples. Products are present in only four lanes; 3, 6, 9, containing the *H. perforatum* samples and lane 10 *H. delphicum*. Numbered lanes in both figures are as follows: 1 *H. androsaemum*, 2 *H. kouytchense*, 3 *H. perforatum*, 4 *H. maculatum*, 5 *H. patulum*, 6 *H. perforatum*, 7 *H. athoum*, 8 *H. calycinum*, 9 *H. perforatum*, 10 *H. delphicum*, 11 *H. ascyron*.

Table 2.2 DNA sequence similarity of *Hypericum* species at primer annealing positions

Hypericum species	% Sequence Similarity at primer annealing positions to *H. perforatum.*		
	FO2	HRI-S	RO
H. perforatum	100	100	100
H. kouytchense	69.6	89.6	100
H. calycinum	65.2	89.6	100
H. ascyron	69.6	86.2	100
H. androsaemum	78.2	89.6	95.8
H patulum	69.6	89.6	100
H. maculatum	100	96.6	100
H. athoum	95.7	96.6	100
H. delphicum	95.7	93.1	100

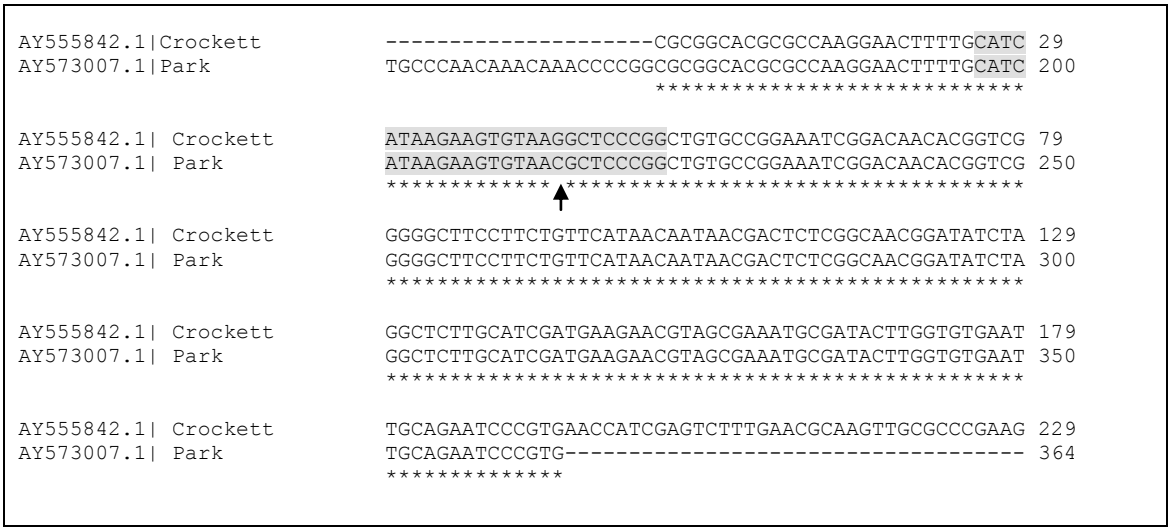


Figure 2.6 Section of the multiple alignment of the two published *H. maculatum* nrITS sequences.

Highlighted by an arrow is the one base difference between the sequences which falls in the annealing site of the primer FO2 (grey). The Crockett sequence is published in Crockett et al. (2004) and Park in Park and Kim (2004)

2.3.3 Commercial Fresh Plant Material

DNA extracted from fresh plant material was then examined in order to ensure the robustness of the assay with non-vouchered plant material. Leaf samples were collected from garden varieties sold as *Hypericum calycinum* and *Hypericum* Hidcote. Further samples were collected from eparate *Hypericum* species grown from commercial seeds.

PCR with primers ITS1 and ITS4 verified the presence of amplifiable DNA within all seven DNA extractions (Figure 2.7). Primers rpoC2 and rpoC4 were also tested, which amplify the candidate plastid barcode region *rpoC*. It has previously been reported that genomic DNA may be more problematic to extract and utilise from plant material than plastid DNA due to low copy number (http://www.kew.org/barcoding/rationale.html). However, both the ITS and *rpoC* regions amplified well, indicating that the extraction procedure had been effective for both nuclear and plastid DNA, and that the PCR was similarly efficient. Of seven *Hypericum* samples tested, a product with the primer pairing FO2 and HRI-S was only found with *H. perforatum* samples (Figure 2.7). This shows that the primer combination can be used to differentiate a number of commercial *Hypericum* species from *H. perforatum*, including the Hidcote cultivar, which is often sold as 'St. John's Wort'

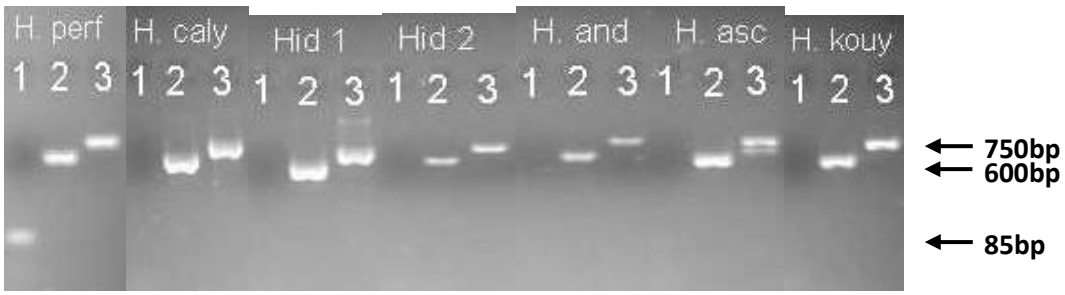


Figure 2.7 Fresh leaf DNA extraction samples with primer pairings.

Primers used are indicated by the numbered lanes: 1 FO2 and HRI-S, 2 *rpoC* 2 and 4, 3 ITS1 and ITS4. All DNA samples are amplified with both *rpoC* 2 and 4 and ITS1 and 4, only the *H. perforatum* sample is amplified with FO2 and HRI-S. Template DNA was as follows: H. perf – *H. perforatum*, H. caly - *H. calycinum*, Hid 1 - Hidcote sample 1, Hid 2 – Hidcote sample 2, H.and – *H. androsaemum*, H. asc – *H. ascyron* and H. kouy – *H. kouytchense*.

2.3.4 Consumer Retail Herbal Medicinal Products

DNA extracted from consumer retail medicinal product samples was subjected to identical PCR conditions and primer pairings as the vouchered samples, with the addition of the primer combination of FO2 and RO. The three products sampled were sold as capsules filled with dried, powdered plant material. Samples from companies A and B gave no products with the primers ITS1 and ITS4 (Figure 2.8). This could be the result of DNA shearing during sample processing. The nrITS region is 750 bp long, as compared to the product of primer pair FO2 & RO which is 160bp and of FO2 & HRI-S which is 85bp. This difference in length shows the impact of DNA quality, since highly processed or aged material is likely to contain sheared DNA. Random cleavage of the DNA between primer annealing positions prevents the formation of the expected product. This is the likely cause of the negative results for nrITS amplification, as short amplicon products were formed using DNA from samples A and B as the template (Figure 2.8). The amount of product from these samples is similar to that of the *H. perforatum* control with the FO2 and HRI-S primers which indicates the presence of *H. perforatum* as stated on the product labels.

The sample from company C gave a product with the generic ITS1 and ITS4 primers, but not with the other two primer pairings (Figure 2.8). This shows that the DNA extraction process was successful, and indicates that amplifiable nuclear DNA was present in the sample, though it was not from *H. perforatum*. This can be explained by the fact that the product is sold as an "Herbal Complex", so is expected to contain material from other plants which will amplify with the generic ITS1 and ITS4 primers. The size of the nrITS product from company C appears to be smaller than that from the *H. perforatum* control, having run slightly further on the gel,

indicating that the sequence between the primers is shorter in this DNA than in *H. perforatum*. However, the label states that 333 mg of *H. perforatum* is present in each capsule, (compared to 116mg of other plant material), so a product would be expected with this primer pairing from both the *H. perforatum* DNA and that from the other plant material. With the pairings of FO2 & RO and FO2 & HRI-S, the company C sample gave no product.

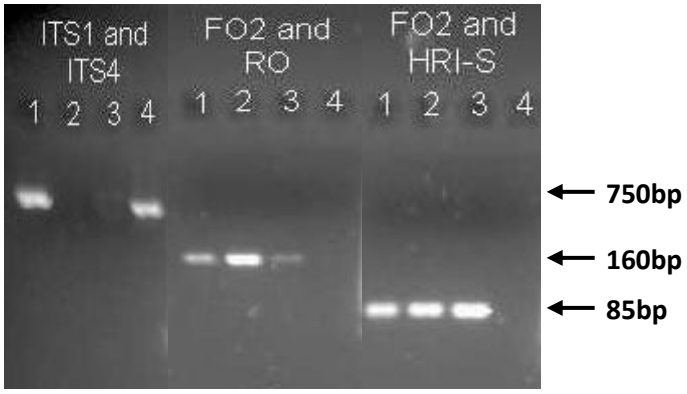


Figure 2.8 PCR products of primer pairings ITS1 and 4, FO2 and RO and FO2 and HRI-S as indicated.

Numbered lanes are as follows: 1 *H. perforatum* positive control, 2 Company A, 3 Company B, 4 Company C. The *H. perforatum* sample yields product with all 3 primer combinations. Samples from Companies A and B do not yield an entire ITS sequence, (primers ITS1 and 4), possibly due to shearing of the DNA. However, as the length of the amplicon shortens and specificity increases a product is seen for both samples indicating the presence of *H. perforatum* DNA in the samples. The sample from Company C yields product with only the generic ITS primers, indicating that the ITS sequence present is not from *H perforatum*.

2.3.5 Assay Sensitivity

The sensitivity of the reaction was then addressed. To replicate an "Herbal Complex" of the type produced by company C, a sample of Ramie (*Boehmeria nivea)* amplifiable nuclear DNA (known not to be amplified with the primer combination of FO2 and HRI-S) was used as a diluent for *H. perforatum* DNA. With an increase from 30 to 40 cycles on the PCR programme used previously, *H. perforatum* was detectable when present as just 0.1% of the total DNA present (Figure 2.9), equivalent to 0.75ng DNA. At this level of amplification, the product remained singular and specific. The sample from company C was also tested with the increased number of cycles, and again no product was seen. Based on the stated composition of the capsules, a product would have been expected from the DNA extracted from company C's St. John's Wort Herbal Complex if it had indeed contained 333mg of *H. perforatum*.

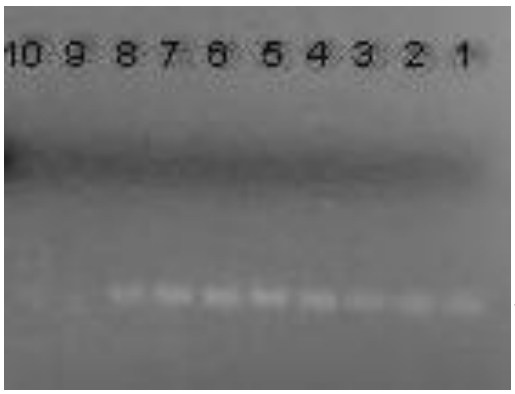


Figure 2.9 Gel electrophoresis of detection level assay.

Lanes 1 – 8 show products from a combination of Ramie, *Boehmeria nivea*, DNA with *H. perforatum* DNA and primers FO2 and HRI-S. *H perforatum* DNA as a % of total DNA present: 0.1% lane 1, 0.2% lane 2, 0.3% lane 3, 0.4% lane 4, 0.5% lane 5, 0.6% lane 6, 0.8% lane 7 and 1% lane 8. Lane 9 is the negative control and Lane 10 DNA from the Company C sample.

2.3.6 Conclusions

2.3.6.1 nrITS

Contrary to the published problems encountered with the nrITS region (variation and alignment difficulties) (Sass et al., 2007), this research found in agreement with Crockett et al. (2004) that the nrITS is an effective region for use within the *Hypericum* genus. This region was readily amplified and the published DNA sequences aligned for all of the *Hypericum* species tested, and is a suitable target for the design of DNA-based assays.

The nrITS region showed high conservation within *Hypericum* species, causing identification of microcodes and primer design to be difficult, but not impossible. The constraints of primer design caused by this sequence conservation are shown by the non ideal features of the primers, such as mononucleotide repeats and amplicons both too long and too short for ideal assay conditions (See section 2.3.1). Despite this, the nrITS region showed sufficient variation for the design of primers to distinguish *H. perforatum* from other closely related species, as was successfully shown. Indeed, as has been reported for the ITS2 region in Chinese medicinal plants (Chen et al., 2010), the nrITS could prove to be the most suitable barcode region for *Hypericum* species (See section 5.3.5.4).

2.3.6.2 Primer Design to Microcodes

The method of designing PCR primers to microcodes within barcoding regions, demonstrated in this study, has enabled the design of a rapid and reliable molecular identification method for *H. perforatum*. This has the potential to become a model for the design of molecular

identification methods, and may be reproducible in other economically valuable plants as barcode data become available.

The resultant PCR test produced is similar to the end product of many of the different DNA-based techniques described in section 1.3, but the method of design differs significantly. In comparison to methods such as RFLP, AFLP and RAPD the design of primers to microcodes is cheaper and less time consuming. The identification of the microcode itself requires careful consideration, but is aided considerably by dedicated software such as Fingerprint (See section 2.3.1).

2.3.6.3 Cross Amplification

The cross amplification of the *H. delphicum* sample is a product of the extremely high sequence conservation found in the nrITS region between *Hypericum* species. This species is endemic to Greece and is not widely distributed, making it unlikely to be found as an adulterant or contaminant of *H. perforatum* products (Crockett et al., 2007). However, it is possible that other *Hypericum* species which have not yet been sequenced or tested with this assay may cross amplify (See section 5.1.2). In order to verify the technique, a wider range of species and samples should be analysed for cross reactions, both within the *Hypericum* genus and across a wide range of other genera.

The use of barcode data as a platform for primer design should reduce the possibility of cross amplification, as these regions will have already been shown to species specific when the sequence of the entire region is used. As multiple regions have been proposed as candidate barcodes, it may be possible to find different microcodes for species in each of these regions, developing a multiple marker approach including each region to further guard against cross amplification causing misidentification.

2.3.6.4 Applicability of DNA based techniques

The DNA degradation found within the commercial samples was found to be mainly shearing, rather than overall degradation which would have prevented the application of DNA based techniques. The attention taken in experimental design to produce an assay with a short amplicon to ensure that sheared DNA would be amplified was justified, based on the results shown in Figure 2.8. The highly processed samples from companies A and B did not produce the full length ITS amplicon (750bp), and particularly for company B even the 160bp amplicon of FO2 and RO is only faintly visible on the gel. However, for both of these samples the 85bp FO2 and HRI-S amplicon was produced, and had a similar intensity on the gel as the positive

control. This suggests that 85bp is a good size of amplicon for use in these circumstances, and also that the nrITS may be too long to use as a comparative measure to this.

The commercial samples tested in this research are an example of a situation in which traditional botanical identification methods are not sufficient. The physical characteristics of the plant are no longer present, rendering the macro and microscopic methods ineffective. Chemical content may be measured, and certainly hypericin and hyperforin content are stated on the labels of each product. However, the chemical content of the entire plant is considered to be the active substance (See section 1.4.2), and this cannot be proven to be present by measuring hypericin and hyperforin. In addition, other *Hypericum* species produce these two constituents, so they cannot be used to verify the plant species present. While chemical content must always be measured in medicinal plant products, DNA-based methods present an ideal complement to these techniques and are capable of identifying species where no other method can.

2.3.6.5 Further Investigation

Further work continued in this area, the consumer retail samples described in section 2.3.4 were tested again from DNA extraction through to the final assay and the same results were found. A survey of the available types of SJW product on sale was also commenced, with a view to further testing of the market. The commercial fresh plant material was also sequenced, with BLAST searches of their nrITS sequences indicating that they were in fact sold as the correct species.

This research has found that one in three SJW consumer retail products is dubious; this is an extremely high proportion and is likely to be an overestimate of the real number of questionable products on the market. Increased sampling of the products available should be conducted to find the real rate of occurrence of dubious products claiming to contain SJW plant material. This is likely to become more difficult as different types of product are selected for testing, in particular tinctures and extracts, discussed further in section 6.3.

This chapter shows that an uncomplicated molecular technique can be used in circumstances where other identification methods are not applicable. The simplicity of the method increases the practicality of its use in industrial applications where DNA sequencing or other more complicated techniques are not realistic.

The research described in this chapter was published in the journal Planta Medica as an Original Paper (Howard et al., 2009).

3 Measuring Medicinal Plant DNA by Quantitative PCR (qPCR)

3.1 Introduction

3.1.1 Medicinal Plant Purity

A quick and simple PCR based identification technique for *H. perforatum* has been developed (See Chapter 2). However, this method is capable only of a positive or negative result; *H. perforatum* either is or is not present. Instances where a different plant has been mixed with *H. perforatum*, either intentionally or inadvertently, would still give a positive result. In these circumstances a quantitative measure would be beneficial to give an indication of how much of the material in question is *H. perforatum*. This is of particular importance to the medicinal plant industry, as proof of the identity of the starting material is only one of the requirements of the European Pharmacopoeia (European Pharmacopoeia, 2008). Purity is an extremely high priority: for instance, SJW must be 98% pure. The conventional PCR assay is not capable of giving a measure of quantity, since a crude estimation based on the size and intensity of the band formed by a specific amplicon is the limit of this essentially qualitative measure.

3.1.2 Quantitative PCR (qPCR)

Quantitative Polymerase Chain Reaction, qPCR, was first described by Higuchi et al. (1993) after having accidentally incorporated ethidium bromide into a PCR reaction. Having realised that the reaction was not inhibited by the presence of ethidium bromide, it was concluded that the reaction could be monitored as it took place, in real-time (Higuchi et al., 1993). This would then result in a measure of input DNA. After each PCR cycle, a fluorescent reading is taken which measures DNA content by virtue of a DNA binding dye present in the reaction. As the reaction takes place, the amount of DNA measured increases, resulting in a trace of the sigmoidal curve of DNA production when plotting cycle number against fluorescence. The point at which this curve crosses the 'threshold' is termed the quantification cycle (Cq) (Bustin et al., 2009), indicated in Figure 3.1. This is where the increase in fluorescence becomes exponential and moves beyond the starting fluorescent value. Using standards of known DNA concentration, a calibration curve can be created of Cq against the log of the DNA concentration. This curve can then be used to quantify the amount of DNA in a given sample.

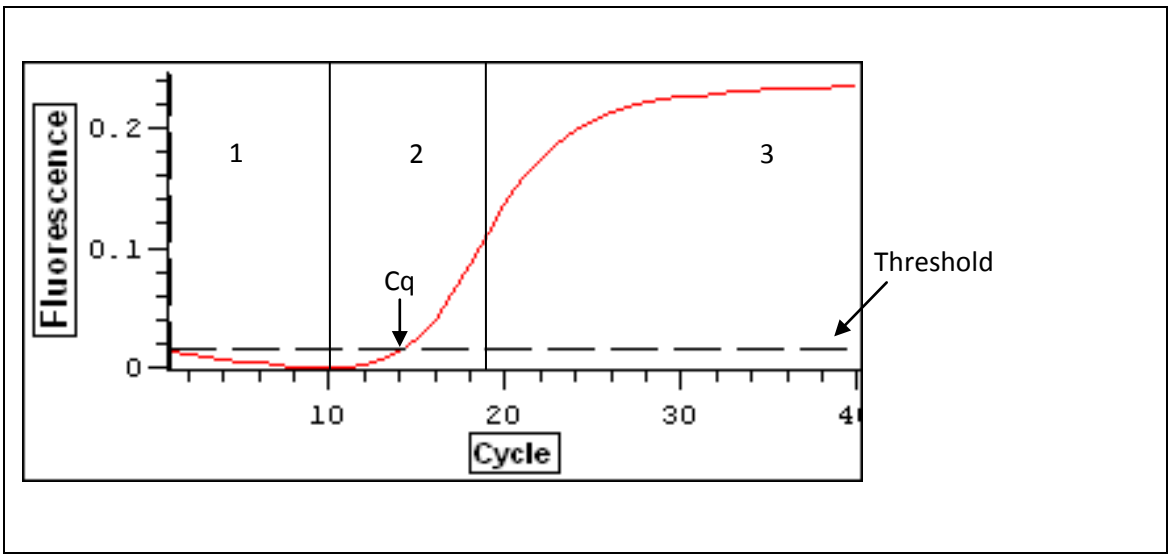


Figure 3.1 Sigmoidal curve of fluorescence increase over the course of a qPCR reaction caused by increasing DNA accumulation.

The regions of the curve indicated are as follows: Section 1 is before duplication has become exponential, showng background fluorescence 'noise'; Section 2 shows the exponential reaction, this region shows the fluorescence measure crossing the threshold where Cq is calculated; Section 3 shows the phase where the qPCR reaction is ceased due to limiting factors such as primer and dNTP availability, causing a plateau in fluorescence. Adapted from (Rebrikov and Trofimov, 2006).

This technique is widely used for many different applications, as is indicated by the number of citations recorded in the ISI Web of Knowledge for the first paper describing the method (Higuchi et al., 1993), 778 (as of 03/04/2010). The attractions of qPCR include the theoretical limit of detection (as PCR requires very few copies of the target DNA sequence to be present for amplification to occur), the specificity (conferred by the primer sequence, annealing temperature and general PCR stringency), and the relative simplicity of application (Bustin et al., 2009). The number and variety of qPCR applications has resulted in the publication of Minimum Information for Publication of Quantitative Real-Time PCR Experiments, the MIQE Guidelines (Bustin et al., 2009). These set out the required information necessary for the publication of qPCR experiments in order for reviewers to adequately assess the assay described.

Table 3.1 MIQE Checklist of information required for publication of qPCR experiments. Items are divided into experimental phases, and their importance either E – Essential or D – Desirable. From (Bustin et al., 2009).

ITEM TO CHECK	IMPORTANCE
EXPERIMENTAL DESIGN	
Definition of experimental and control groups	**E**
Number within each group	**E**
Assay carried out by core lab or investigator's lab?	D
Acknowledgement of authors' contributions	D
SAMPLE	
Description	**E**
Volume/mass of sample processed	D
Microdissection or macrodissection	**E**
Processing procedure	**E**
If frozen - how and how quickly?	**E**
If fixed - with what, how quickly?	**E**
Sample storage conditions and duration (especially for FFPE samples)	**E**
NUCLEIC ACID EXTRACTION	
Procedure and/or instrumentation	**E**
Name of kit and details of any modifications	**E**
Source of additional reagents used	D
Details of DNase or RNAse treatment	**E**
Contamination assessment (DNA or RNA)	**E**
Nucleic acid quantification	**E**
Instrument and method	**E**
Purity (A260/A280)	D
Yield	D
RNA integrity method/instrument	**E**
RIN/RQI or Cq of 3' and 5' transcripts	**E**
Electrophoresis traces	D
Inhibition testing (Cq dilutions, spike or other)	**E**
REVERSE TRANSCRIPTION	
Complete reaction conditions	**E**
Amount of RNA and reaction volume	**E**
Priming oligonucleotide (if using GSP) and concentration	**E**
Reverse transcriptase and concentration	**E**
Temperature and time	**E**
Manufacturer of reagents and catalogue numbers	D
Cqs with and without RT	D*
Storage conditions of cDNA	D
qPCR TARGET INFORMATION	
If multiplex, efficiency and LOD of each assay.	**E**
Sequence accession number	**E**
Location of amplicon	D
Amplicon length	**E**
In silico specificity screen (BLAST, etc)	**E**
Pseudogenes, retropseudogenes or other homologs?	D
Sequence alignment	D
Secondary structure analysis of amplicon	D
Location of each primer by exon or intron (if applicable)	**E**
What splice variants are targeted?	**E**
qPCR OLIGONUCLEOTIDES	
Primer sequences	**E**
RTPrimerDB Identification Number	D
Probe sequences	D**
Location and identity of any modifications	**E**
Manufacturer of oligonucleotides	D
Purification method	D
qPCR PROTOCOL	
Complete reaction conditions	**E**
Reaction volume and amount of cDNA/DNA	**E**
Primer, (probe), Mg++ and dNTP concentrations	**E**
Polymerase identity and concentration	**E**
Buffer/kit identity and manufacturer	**E**
Exact chemical constitution of the buffer	D
Additives (SYBR Green I, DMSO, etc.)	**E**
Manufacturer of plates/tubes and catalog number	D
Complete thermocycling parameters	**E**
Reaction setup (manual/robotic)	D
Manufacturer of qPCR instrument	**E**
qPCR VALIDATION	
Evidence of optimisation (from gradients)	D
Specificity (gel, sequence, melt, or digest)	**E**
For SYBR Green I, Cq of the NTC	**E**
Standard curves with slope and y-intercept	**E**
PCR efficiency calculated from slope	**E**
Confidence interval for PCR efficiency or standard error	D
r2 of standard curve	**E**
Linear dynamic range	**E**
Cq variation at lower limit	**E**
Confidence intervals throughout range	D
Evidence for limit of detection	**E**
If multiplex, efficiency and LOD of each assay.	**E**
DATA ANALYSIS	
qPCR analysis program (source, version)	**E**
Cq method determination	**E**
Outlier identification and disposition	**E**
Results of NTCs	**E**
Justification of number and choice of reference genes	**E**
Description of normalisation method	**E**
Number and concordance of biological replicates	D
Number and stage (RT or qPCR) of technical replicates	**E**
Repeatability (intra-assay variation)	E
Reproducibility (inter-assay variation, %CV)	D
Power analysis	D
Statistical methods for result significance	**E**
Software (source, version)	E
Cq or raw data submission using RDML	**D**

3.1.3 Modifications

The popularity of qPCR and the number of applicable fields has resulted in many types of modification of the original method. Most divergent are the probes which report amplification occurrence, the range having benefited from innovations in fluorescent dyes and quenchers. Three of the most popular methods are TaqMan probes (Livak et al., 1995), Molecular Beacons (Tyagi and Kramer, 1996) and DNA intercalating dyes.

TaqMan probes (Livak et al., 1995) are oligonucleotides designed to anneal initially to the target region of DNA and, as the reaction progresses, to the amplification products of the PCR reaction. Attached to the 5′ end of the probe is a fluorescent dye, and on the 3′ end a quencher which prevents detection of the fluorescent dye due to its proximity. As the reaction takes place, the *Taq* enzyme cleaves the dye from the 5′ end of the probe in the process of replicating the template to which it was bound. This distances the dye from the quencher, enabling the energy produced by its excitation to be recorded and monitored. As more PCR products are formed, more TaqMan probes are hydrolysed giving a method by which to determine the number of amplicons produced (Figure 3.2).

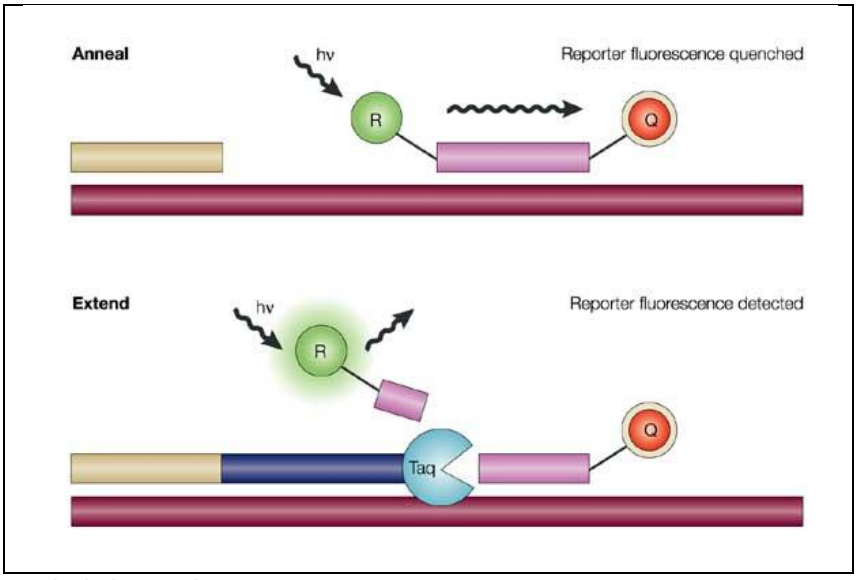


Figure 3.2 TaqMan hydrolysis probe.

The action of the *Taq* enzyme distances the reporter fluorophore (R) from the quencher (Q). Fluorescence emitted when excited by an external light source (hv) at each PCR cycle is proportional to the amount of product formed. Adapted from (Koch, 2004).

Molecular Beacons (Tyagi and Kramer, 1996) act in a similar way to TaqMan probes, in that a fluorophore is attached to the 5′ end and detection of it is prevented by the presence of a quencher on the 3′ end. In this case, the oligonucleotide is designed to form a hairpin structure, causing both ends to be in close proximity. The middle section of the probe is

designed to anneal to the PCR product, and at each annealing stage, bind to it. This separates the fluorophore from the quencher and enables detection.

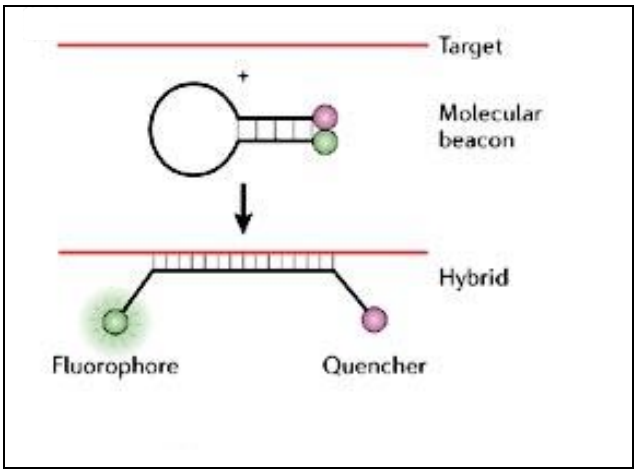


Figure 3.3 Molecular Beacon.

In the absence of the target sequence, the structure forms a stable hairpin preventing detection of the fluorophore. The loop section of the probe is designed to anneal to the target product, distancing the fluorophore from the quencher, and thereby enabling detection which is proportional to the amount of product formed. Adapted from (Condon, 2006).

TaqMan probes and molecular beacons can increase the specificity of the qPCR reactions as they encompass the stringency of the PCR primers and their own sequence to anneal to only the target PCR product.

DNA intercalating dyes, such as SYBR green and ethidium bromide can also be used for qPCR amplicon detection. In the same way that PCR products are visualised on electrophoresis gels, these dyes bind to double stranded DNA which then causes them to fluoresce. In qPCR, these dyes bind to the product and as the number of copies increases so does the fluorescent emission from the dye. This method of detection is much cheaper than methods requiring modified oligonucleotides, and for this reason could be more attractive for routine testing use in industry. However, there are difficulties associated with the use of intercalating dyes, particularly their non-specificity, as any double stranded DNA is detected, including any primer dimers or cross-amplification products. This can be overcome by assay design and analysis of the amplicons produced by either gel electrophoresis or melt curve analysis (Kubista, 2008).

Melt Curve analysis is performed after the qPCR reaction is complete. The temperature is steadily increased and fluorescence monitored at precise temperature increments. Each qPCR amplicon will have a different melting temperature due to its length and sequence, in the

same way that PCR primers have different melting temperatures (*Tm*s). The desired target amplicon is detected by the specific and sharp fall in fluorescence at the melting temperature of the amplicon (82°C in the example shown in Figure 3.4). If a qPCR reaction has produced one specific product, one specific drop in fluorescence is seen. Non-specific amplification produces melt curves with slower declines and multiple dips in fluorescence. Amplicons from the same primer pair should have the same *Tm*.

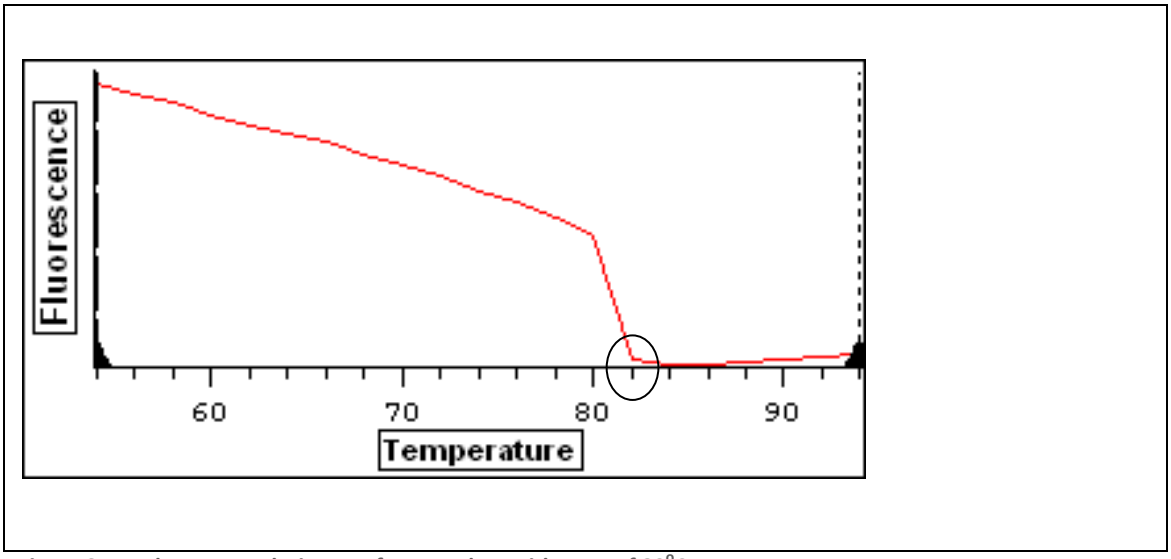


Figure 3.4 Melt curve analysis trace for a product with a *Tm* of 82°C.

Melting temperature is measured by the point at which the fluorescence drops significantly due to the melting of the amplicon. Adapted from (Wilhelm and Pingoud, 2003)

3.1.4 Proposal for Assay Design

The design of this qPCR assay is based on the premise that it will be possible to accurately measure *H. perforatum* DNA by microcode qPCR, and use this measure to assess the purity of a sample. In order to calculate the purity of a sample, two qPCR assays are required; one to measure the DNA of the target species and another to measure the total DNA present. The discrepancy between these two measures would then indicate the presence of another contaminating DNA within the sample.

These assays were designed for SJW based on the nrITS and coding regions. The conserved coding regions were used to design universal primers to amplify total DNA present and the spacers to design species specific primers as described previously (Section 3.2.2).

3.2 Materials and Methods

3.2.1 Sample Material and DNA Extraction

Fresh plant material was cultivated from seeds supplied by Chiltern Seeds Ltd. (Bortree Stile, Ulverston, Cumbria, U.K. LA12 7PB) for the following species: *H. perforatum* (Ref. 701C), *H. kouytchense* (Ref. 700J), *H. androsaemum* (Ref. 698E) and *H. ascyron* (Ref. 698N). Samples were taken after four weeks from seedlings representing each species and subjected to DNA extraction.

DNA extraction was carried out utilising the Qiagen DNeasy Plant Mini Kit (Qiagen Inc., CA) and TissueLyser. Samples were 0.1g fresh leaf material, roughly shredded before disruption. Manufacturer's instructions were followed, with two disruption steps of 1min at 30 Hz with sample, 400µl Buffer AP1 and 4µl RNase A (Further details in section 2.2.3). The DNA was quantified using a Qubit® Fluorometer and Quant-iT™ dsDNA BR Assay Kit (Invitrogen, Carlsbad, CA, USA).

All samples were analysed by sequencing of the nrITS region to confirm species by comparison to published sequences. Several dilution series were made using *H. perforatum* DNA and water, as shown in Table 3.2. A further dilution series of *H. perforatum* in *H. kouytchense* DNA is shown in Table 3.3.

Table 3.2 *H. perforatum* DNA dilution series, three identical series were made separately, named HPD, HD and PD.

Dilution series			*H. perforatum*
HPD	HD	PD	DNA ng/µL
1	1	1	84.40
2	2	2	16.88
3	3	3	3.38
4	4	4	0.68
5	5	5	0.14
6	6	6	0.03
7	7	7	0.01

Table 3.3 Mixed DNA calibration dilution series

PP Dilution series	*H. perforatum* DNA ng/µL	*H. kouytchense* DNA ng/µL	Total DNA ng/µL
1	84.400	0.000	84.40
2	16.880	67.520	84.40
3	3.380	81.020	84.40
4	0.680	83.730	84.40
5	0.140	84.270	84.40
6	0.030	84.370	84.40
7	0.005	84.395	84.40

3.2.2 Primer Design

Published *Hypericum* nrITS DNA sequences were obtained from the GenBank database, with species and Accession Numbers as follows: AY555840 *H. perforatum*, AY555853 *H. kouytchense*, AY555861 *H. calycinum*, AY555849 *H. ascyron*, AY573012 *H. androsaemum*, AY573007 *H. maculatum*, AY555860 *H. patulum*, AY555846 *H. athoum* and AY555845 *H. delphicum*.

A multiple alignment was carried out using the ClustalW program (Chenna et al., 2003) hosted on the European Bioinformatics Institute (EBI) website (www.ebi.ac.uk). All settings were default; some manual gap alignment was conducted afterwards (Figure 2.3).

The specific *H. perforatum* primers used were FO2, FIn, HR373 and HRI-S as previously described (Section 2.2.4, shown in Figure 2.3). The generic primers designed and used were HypGF (5′-CCGTGAACCATCGAGTCTTT-3′) and HypGR (5′-GTCTTACAACCACCGCTGGT-3′), and 2F (5′-ACCAGCGGTGGTTGTAAGAC-3′) and 2R30 (5′- CGAGCAATGCAAGGCTCACG-3′) shown in Figure 3.5. The primer design software used was Primer3 (Rozen S., 2000). The generic primers *rpoC* 2 and *rpoC* 4 (made publicly available on the Royal Botanical Gardens website www.kew.org/barcoding) for the potential barcode *rpoC* plastid region (described in Section 15) were also tested.

3.2.3 qPCR Protocols

PCR reactions, final volume 20µL, in 0.2mL polypropylene reaction tubes with optical caps (Applied Biosystems, ABI, USA) consisted of EXPRESS SYBR® GreenER™ qPCR SuperMix (Invitrogen, Carlsbad, CA, USA) (1x), relevant primers (200nM each), nuclease-free water and template DNA (varies). The MJ Research PTC-200 Peltier Thermal Cycler System and Chromo4™ Detector and Opticon Monitor™ Analysis Software v3.1 (MJ Research, Waltham, MA, USA) were used with the typical program being: 2min at 50^{o}C activation step, 2min at 95^{o}C initial denaturation step, 40 cycles consisting of 15s at 95^{o}C, 1min at 60.1^{o}C, plate read, followed by a melt curve from 54-95^{o}C at a rate of +1^{o}C per 10s, read every 2^{o}C. Fluorescence values are baseline subtracted, baseline measured as global minimum - the lowest fluorescence value measured for each individual sample. The threshold was set at 1 standard deviation above the mean fluorescence values measured over cycles 1-5. Gradient runs were conducted with a range of annealing temperatures shown in Table 3.5, all other parameters were as described above.

Samples without template DNA were utilised as controls. PCR products were run on 3% (w/v) agarose, 0.5 X TBE gels with 2µl SYBRsafe™ (Invitrogen, Carlsbad, CA, USA) DNA stain at 90V for ~30min and analysed in a Bio-Rad Illuminator (Bio-Rad, Richmond, CA, USA) with ChemiDocXRS Camera and Quantity One software to ensure that only the target amplicon was produced.

3.3 Results

3.3.1 Optimisation and Selection of Primers

3.3.1.1 Universal Primers

Three pairs of generic primers: 2F + 2R30, HypGF + HypGR, and rpoC2 + rpoC4, were tested. The primer pair rpoC2 + rpoC4 are designed to one of the previously proposed barcoding regions *rpoC1* (See section 1.3.9). This region was not chosen as a barcode due to low species resolution but was readily amplified in a large number of species. The two primer pairs 2F + 2R30 and HypGF + HypGR were designed for qPCR particularly, using the Primer3 software programme (Rozen S., 2000) and the published *H. perforatum* nrITS sequence. The primers were then checked in a multiple alignment of further published nrITS sequences from *Hypericum* species to ensure that any *Hypericum* DNA would be amplified using the primers (Figure 3.5). As shown in Table 3.4, the pair with the highest sequence similarity was HypGF + HypGR.

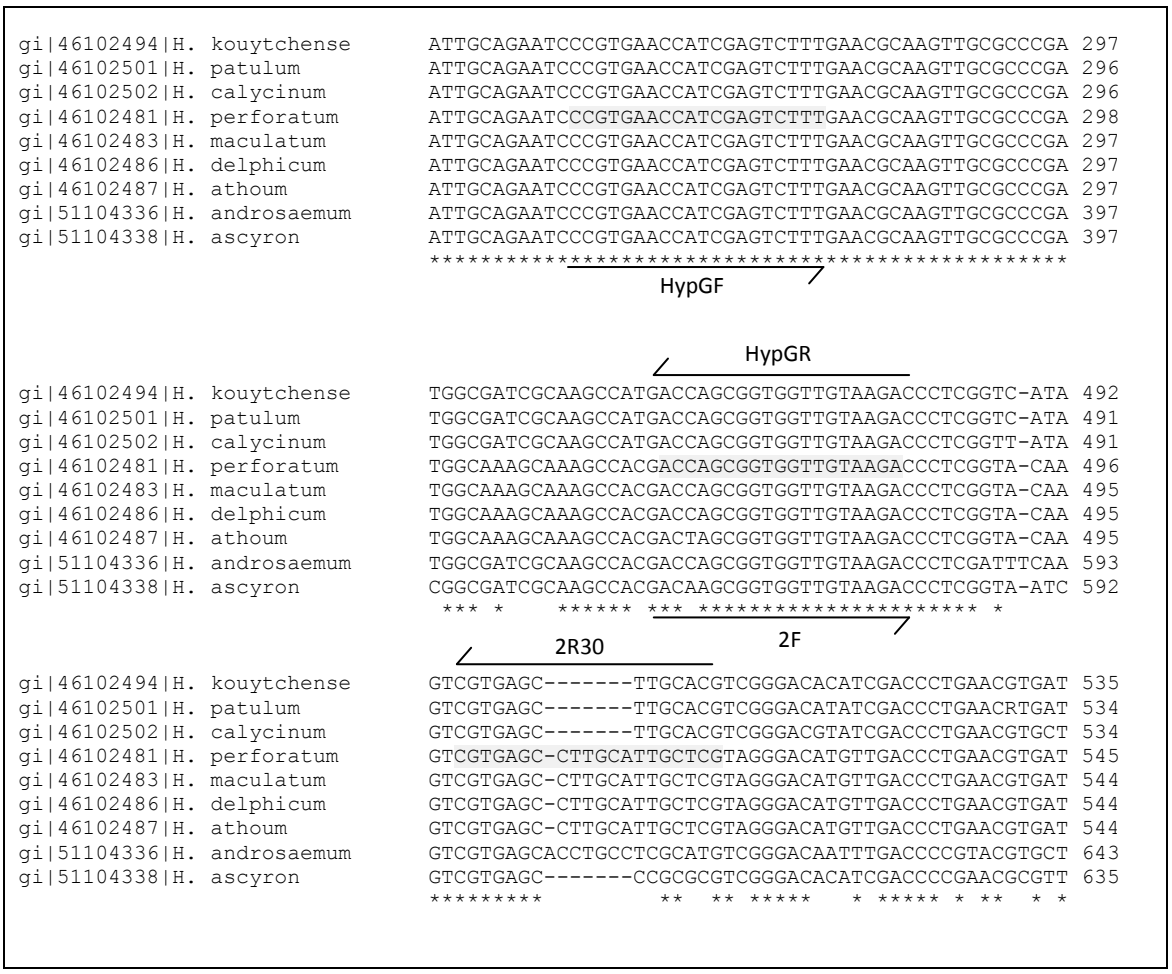


Figure 3.5 Section of a multiple alignment of the nrITS regions of nine *Hypericum* species with qPCR primer annealing positions indicated.

Table 3.4 Sequence similarity, %, at qPCR primer annealing positions of nine *Hypericum* species.

Hypericum species	% Sequence Similarity at primer annealing positions to *H. perforatum.*		
	HypGF	HypGR/2F	2R30
H. perforatum	100	100	100
H. kouytchense	100	100	50
H. calycinum	100	100	50
H. ascyron	100	95	55
H. androsaemum	100	100	75
H patulum	100	100	50
H. maculatum	100	100	100
H. athoum	100	95	100
H. delphicum	100	100	100

All three universal primer pairs were tested for qPCR efficiency using a temperature gradient run: identical qPCR reactions with different annealing temperatures. Although primer melting and annealing temperatures can be calculated using formulae, optimum results are obtained for primer pairs by empirically testing annealing temperatures to find the most suitable (Bustin et al., 2009). For this purpose, thermal cyclers are designed to enable a range of annealing temperatures to be used in one run, a gradient run. Figure 3.6 shows the gradient calculator used to set the different annealing temperatures used throughout a gradient run. The annealing temperature used directly affects the efficiency and Cq of qPCR reactions. Efficiency is calculated here based on the slope of the amplification curve in the log-linear phase of the qPCR reaction, based on the formula $E = 10^{(-1/slope)} - 1$. The log-linear phase is characterised by the number of amplicons increasing exponentially (Figure 3.1). The threshold is also set in this portion, as near as possible to the beginning of the phase.

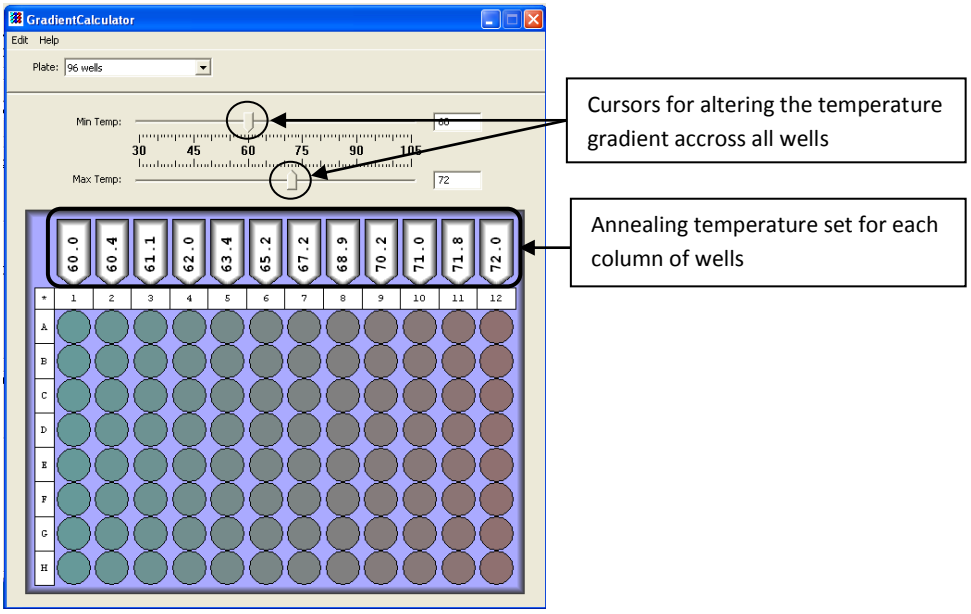


Figure 3.6 Gradient Calculator used with qPCR equipment.

The annealing temperature in the cycling program is altered, all other conditions remain constant. Moving the arrows on the minimum and maximum temperature measures alters the annealing temperatures in each well to the preference of the user. Annealing temperature for each column of wells is indicated.

The gradient run qPCR amplification and melt curves for the three potential universal primer pairs tested are shown in Figure 3.7 to Figure 3.12. Each trace represents a different annealing temperature ranging from 54°C to 63.3°C. The colour key for each temperature is in Table 3.5. All three primer pairs amplified well, producing classic sigmoidal amplification curves. The melt curves for all pairs showed amplicon specificity producing identical *Tm*s.

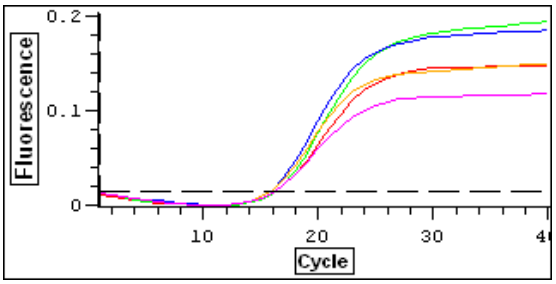


Figure 3.7 Gradient run amplification traces for 2F and 2R30

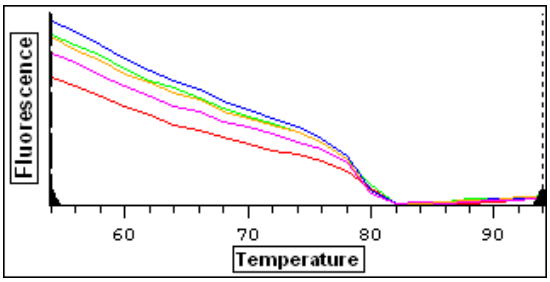


Figure 3.8 Gradient run melt curve traces for 2F and 2R30

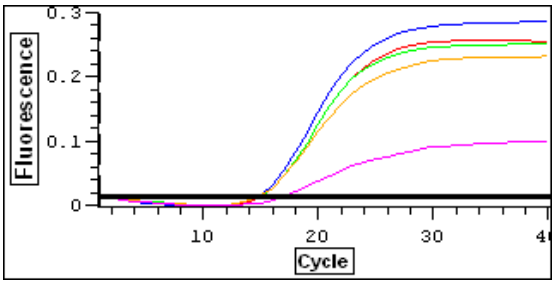


Figure 3.9 Gradient run amplification traces for HypGF and HypGR

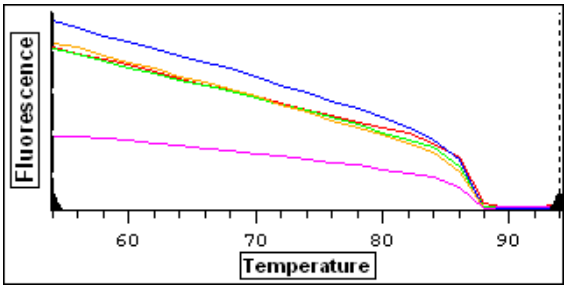


Figure 3.10 Gradient run melt curve traces for HypGF and HypGR

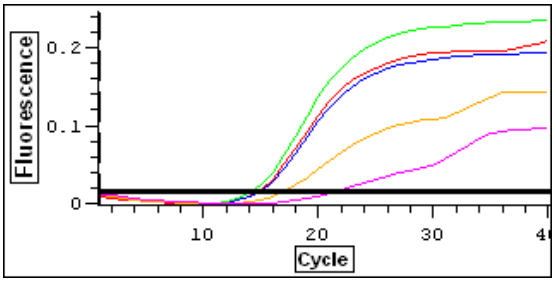


Figure 3.11 Gradient run amplification traces for rpoC 2 and 4

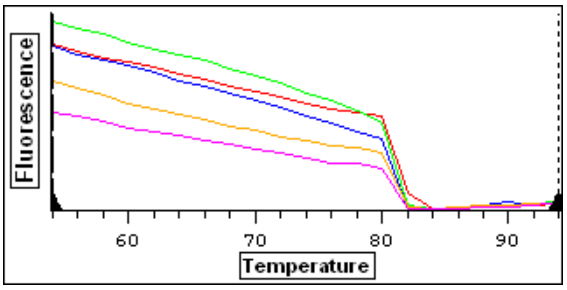


Figure 3.12 Gradient run melt curve traces for rpoC 2 and 4

The results from the gradient run were analysed to determine the optimum annealing temperature for each pair (Table 3.5). The primer pair most susceptible to altering the annealing temperature was rpoC2 and rpoC4 with Cq values ranging from 14.16 to 21.26 and efficiencies from 34.42 to 95.87% (Table 3.5). All three primer pairs achieved the highest efficiencies at the annealing temperature of 57.1°C, with HypGF and HypGR reaching 114.47%, above the theoretical maximum of 100%. Due to the high efficiency scores and low Cq values obtained (Table 3.5), and the high sequence similarity between *Hypericum* species (Table 3.4), the *Hypericum* generic primers HypGF and HypGR were selected for the qPCR assay.

In addition to the aforementioned features, the primers HypGF and HypGR produce a PCR amplicon of just 225bp (Figure 3.21), as compared to several hundred bases for rpoC2 and rpoC4, another pair tested. A smaller amplicon is generally preferable for qPCR (Ambion), but also particularly in the case of DNA samples of unknown quality. Plant material used in medicinal herbal products is often highly processed and/or aged; this can lead to DNA degradation. In these circumstances, shorter amplicons are more reliable as their templates are less likely to have been subject to shearing (See Section 2.3.1). Should the DNA be sheared between the annealing positions of the primers an amplicon will not be formed.

Table 3.5 Gradient run qPCR results for three universal primer pairs tested. The annealing temperature at which the highest efficiency was achieved is highlighted for each pair. All reactions contained equal concentrations of template DNA and primers.

Primer Pair	Figure number	Trace colour on graph	Annealing Temperature, °C	Efficiency, %	Cq	Melt Curve Quality
2F and 2R30	Amplification - Figure 3.7 Melt Curve - Figure 3.8	Red	54.0	67.69	16.24	Good
		Green	55.0	84.22	16.26	
		Blue	57.1	94.82	15.81	
		Orange	60.6	88.83	15.84	
		Pink	63.3	67.22	16.21	
HypGF and HypGR	Amplification - Figure 3.9 Melt Curve - Figure 3.10	Red	54.0	85.39	15.01	Very Good
		Green	55.0	86.32	15.02	
		Blue	57.1	114.47	14.75	
		Orange	60.6	109.18	14.94	
		Pink	63.3	52.34	17.13	
rpoC 2 and 4	Amplification - Figure 3.11 Melt Curve - Figure 3.12	Red	54.0	91.36	14.57	Good
		Green	55.0	84.23	14.16	
		Blue	57.1	95.87	14.88	
		Orange	60.6	70.84	16.98	
		Pink	63.3	34.42	21.26	

3.3.1.2 Specific Primers

The four pairs of *H. perforatum* specific primers described in section 2.3.2 were tested for qPCR suitability in the same way as the universal pairs. The gradient run qPCR amplification and melt curves for the four specific primer pairs tested are shown below (Figure 3.13 to Figure 3.20). Each trace represents a different annealing temperature ranging from 54°C to 63.3°C, with the colour key shown in Table 3.6. The primer pair with the highest recorded efficiency is FIn + HR373. However, this primer pair also has the largest range of efficiency scores, from 72.35% to 110.35%, showing that it is the most susceptible to alteration of the annealing temperature. The relationship between qPCR and annealing temperature should be such that a peak of efficiency is seen at the optimum temperature, and either side of this the temperature the efficiency decreases, as shown with FO2 + HRI-S in Table 3.6. The pair FIn + HR373 show two seperate peaks in efficiency, possibly due to non-specific amplification as

indicated in the melt curve for this reaction (Figure 3.14). Due to this, the FIn + HR373 pairing was not selected for the qPCR assay. The other three primer pairs all produced normal amplification and melt curves.

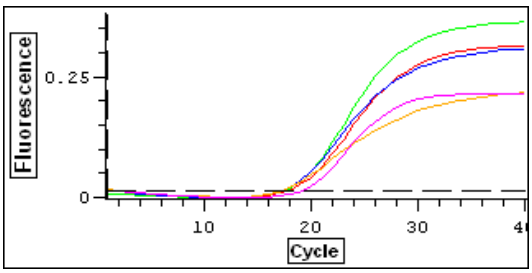


Figure 3.13 Gradient run amplification traces for FIn and HR373.

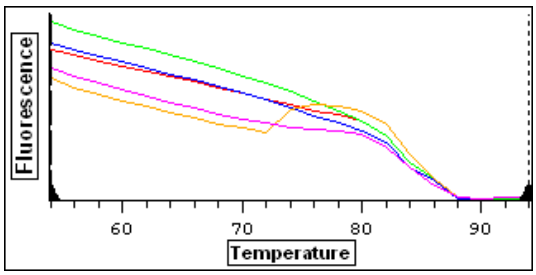


Figure 3.14 Gradient run melt curve traces for FIn and HR373.

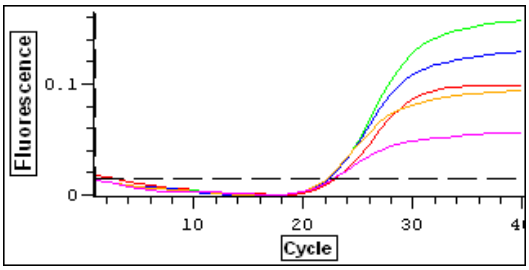


Figure 3.15 Gradient run amplification traces for FIn and HRI-S.

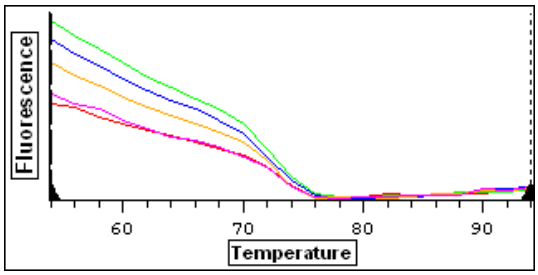


Figure 3.16 Gradient run melt curve traces for FIn and HRI-S.

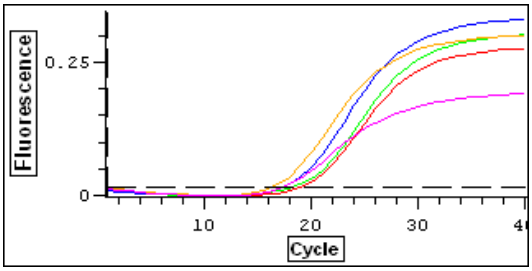


Figure 3.17 Gradient run amplification traces for FO2 and HR373.

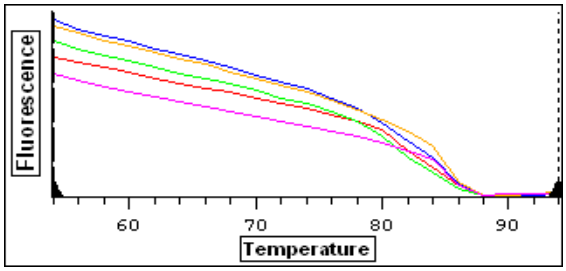


Figure 3.18 Gradient run melt curve traces for FO2 and HR373.

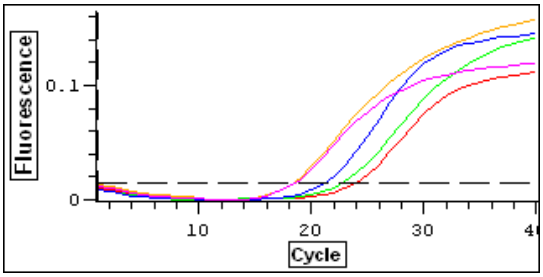


Figure 3.19 Gradient run amplification traces for FO2 and HRI-S.

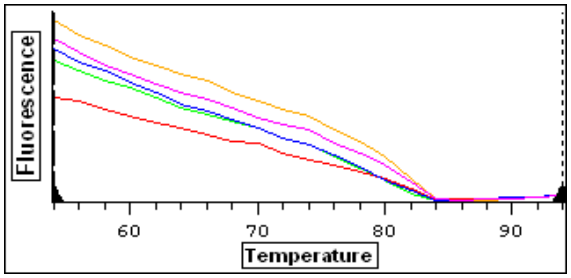


Figure 3.20 Gradient run melt curve traces for FO2 and HRI-S.

Table 3.6 Gradient run qPCR results for four specific primer pairs tested. The annealing temperature at which the highest efficiency was achieved is highlighted for each pair. All reactions contained equal concentrations of template DNA and primers.

Primer Pair	Figure number	Trace colour on graph	Annealing Temperature, °C	Efficiency, %	Cq	Melt Curve Quality
FIn and HR373	Amplification - Figure 3.13 Melt Curve - Figure 3.14	Red	54.0	72.35	18.00	Low
		Green	55.0	83.11	17.53	
		Blue	57.1	78.65	17.37	
		Orange	60.6	75.86	17.37	
		Pink	63.3	110.35	19.26	
FIn and HRI-S	Amplification - Figure 3.15 Melt Curve - Figure 3.16	Red	54.0	58.41	23.00	Good
		Green	55.0	79.67	22.36	
		Blue	57.1	70.76	22.36	
		Orange	60.6	66.71	22.14	
		Pink	63.3	60.10	22.84	
FO2 and HR373	Amplification - Figure 3.18 Melt Curve - Figure 3.19	Red	54.0	76.71	18.95	Good
		Green	55.0	59.85	18.27	
		Blue	57.1	80.70	17.38	
		Orange	60.6	79.55	16.28	
		Pink	63.3	73.33	17.36	
FO2 and HRI-S	Amplification - Figure 3.19 Melt Curve - Figure 3.20	Red	54.0	53.71	23.91	Good
		Green	55.0	49.40	22.71	
		Blue	57.1	56.10	21.25	
		Orange	60.6	59.06	18.31	
		Pink	63.3	52.54	18.46	

The pair FIn + HRI-S was not selected for two main reasons. Firstly, the primer FIn has a 70% GC content (Table 3.7) and also a run of five consecutive Gs close to the 3′ end, neither of which attributes is recommended for qPCR (PremierBiosoft, 2010). Secondly, the annealing positions of these two primers overlap by two base pairs.

The remaining two pairs were FO2 + HRI-S, and FO2 + HR373. The pair which amplified with the highest efficiency was FO2 + HR373, Table 3.6. However the amplicon produced by FO2 +HR373 is 293 bp in length, above the recommended length for qPCR. FO2 and HRI-S have identical *Tm*s (Table 3.7), have previously been shown to be specific to *H. perforatum* (Howard et al., 2009) and produce an ideally sized amplicon of just 85bp. For these reasons they were selected for the qPCR assay.

Table 3.7 *H. perforatum* specific primer attributes.

The *Tm* is initially calculated using +2°C A/T and +4°C G/C, the *Tm* with length calculation corrects for primers longer than 13 nucleotides; (64.9+41)*((number GC-16.4)/length) from (Wallace et al., 1979).

Primer	Length, bases	% GC Content	*Tm*, °C	*Tm* with length calculation, °C
FIn	20	70	68	60
FO2	23	52	70	57
HR373	24	46	70	56
HRI-S	29	38	80	57

3.3.1.3 Primer pairs HypGF + HypGR, and FO2 + HRI-S

Figure 3.21 shows the annealing position of the specific and universal primer pairs; the HRI-S annealing position terminates only 75bp from the beginning of the HypGF annealing position. As a result, the generic primers act as both a measure of the total DNA content in a sample and a verification of the presence of the nrITS region within the DNA sample, i.e. a positive control. As is the case with short amplicons, annealing positions sufficiently close to one another give a measure of security against DNA shearing. Should neither pair produce an amplicon then it may be concluded that this region of DNA is not present or not of sufficient quality to be measured via qPCR.

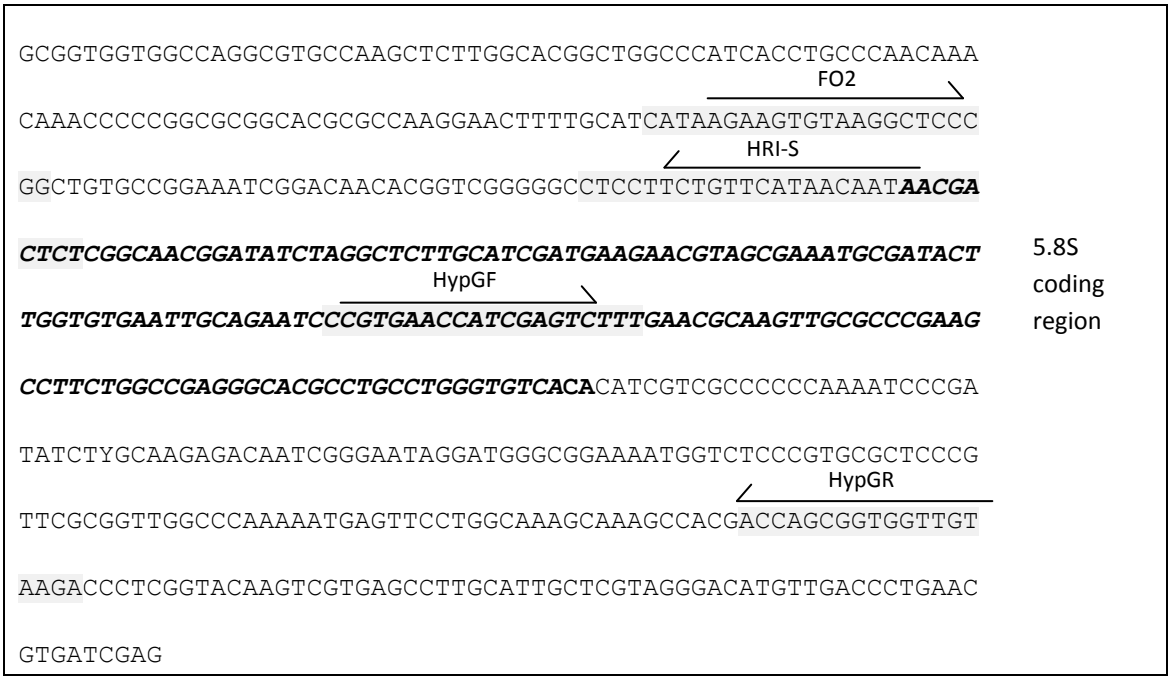


Figure 3.21 DNA sequence of *H. perforatum* nrITS with qPCR primer annealing positions shown in blue, named above the DNA sequence. The highly conserved 5.8S coding region is indicated in italics.

3.3.2 *H. perforatum* DNA dilutions

Three dilution series of genomic *H. perforatum* DNA in water (Table 3.2) were used to calibrate both primer pairs in qPCR. The series covered a wide range, from 0.0054ng/µL to 84ng/µL, ensuring coverage of the possible yields obtained from DNA extraction methods. The Qiagen DNeasy Plant Mini Kit states a maximum genomic DNA yield of 150ng/µL (Qiagen, 2006) from fresh plant material. However, as highly processed and somewhat aged plant material is used for medicinal herbal products, both lower yield and lower quality DNA was expected. In addition to this, the medicinal plant material subjected to DNA extraction may not be pure, whether intentionally sold as a mixture of plants or inadvertently adulterated with misidentified plant material. This results in the target plant DNA making up an unknown percentage of the total DNA extracted, potentially much less than 150ng/µL.

The replicate dilution series were made and measured with both the generic and specific primer pairs. An example of the results from HypGF + HypGR is shown in Figure 3.22. Both the specific and generic primer pairs yielded highly reproducible results, as is revealed by the very low coefficient of variance values, ranging from 0.01 to 0.03 for FO2 + HRI-S (Table 3.8) and 0.01 to 0.05 for HypGF + HypGR (Table 3.9). All of the results obtained are shown in Figure 3.23 and Figure 3.24.

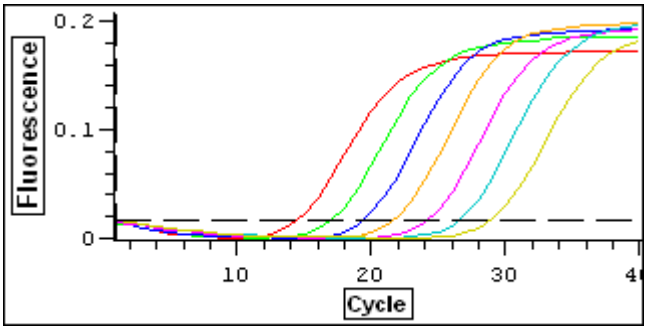


Figure 3.22 Dilution series traces for the generic primer pair, HypGF and HypGR.

Each trace represents a different input DNA concentration (ng/µL): Red – 84.4, Green – 16.88, Dark Blue – 3.38, Orange – 0.68, Pink – 0.14, Light Blue – 0.03 and Yellow – 0.005.

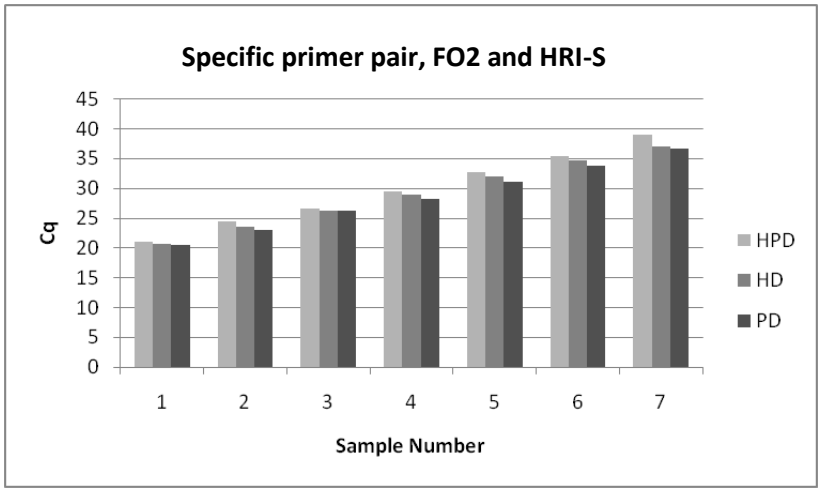


Figure 3.23 Bar graph showing the Cq values obtained from the specific primer pair, FO2 and HRI-S, with each sample within the three dilution series described in Table 3.2.

Table 3.8 Standard Deviation, Coefficient of Variance and mean of Cq values obtained across three dilution series' described in Table 3.2 with the specific primers FO2 and HRI-S

Sample	Mean	Standard Deviation	Coefficient of Variance
1	20.72	0.25	0.01
2	23.71	0.71	0.03
3	26.37	0.27	0.01
4	28.95	0.60	0.02
5	31.96	0.77	0.02
6	34.66	0.83	0.02
7	37.59	1.25	0.03

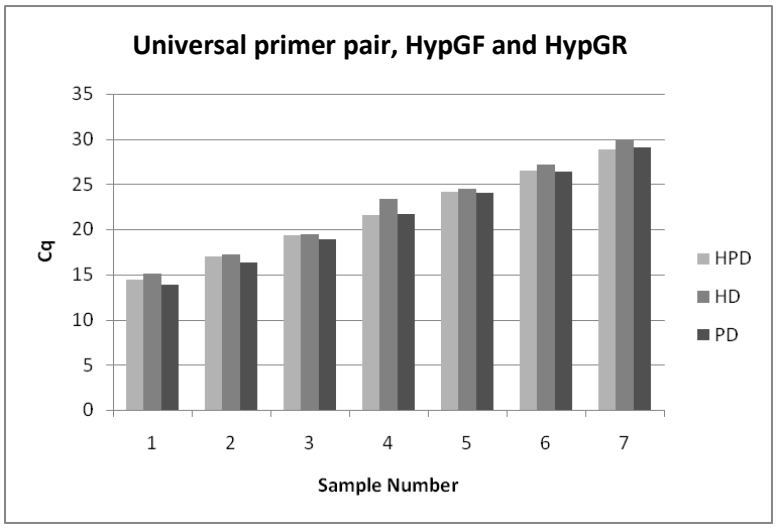


Figure 3.24 Bar graph showing the Cq values obtained from HypGF and HypGR with each sample within the three dilution series described in Table 3.2.

Table 3.9 Standard Deviation, Coefficient of Variance and mean of Cq values obtained across the dilution series' listed in Table 3.2 with the universal primers HypGF and HypGR.

Sample	Mean	Standard Deviation	Coefficient of Variance
1	14.46	0.64	0.04
2	16.87	0.47	0.03
3	19.26	0.33	0.02
4	22.21	1.01	0.05
5	24.24	0.24	0.01
6	26.66	0.44	0.02
7	29.25	0.50	0.02

The mean Cq values for each data point were plotted against the log of the DNA concentration of each sample. The generic primers HypGF and HypGR gave a calibration graph with an R^2 value of 0.999 (Figure 3.25). The specific primers FO2 and HRI-S gave a calibration with an R^2 value of 0.999 (Figure 3.26). The actual Cq values for the two primer pairs at each data point differed due to the efficiency of each primer combination. The specific pairing is much less efficient than the generic due to the compromise required to balance efficiency and specificity. To ensure amplification of only *H. perforatum* DNA, sequence regions which did not contain ideal features for primer design had to be included, resulting in specific primers with low efficiency scores.

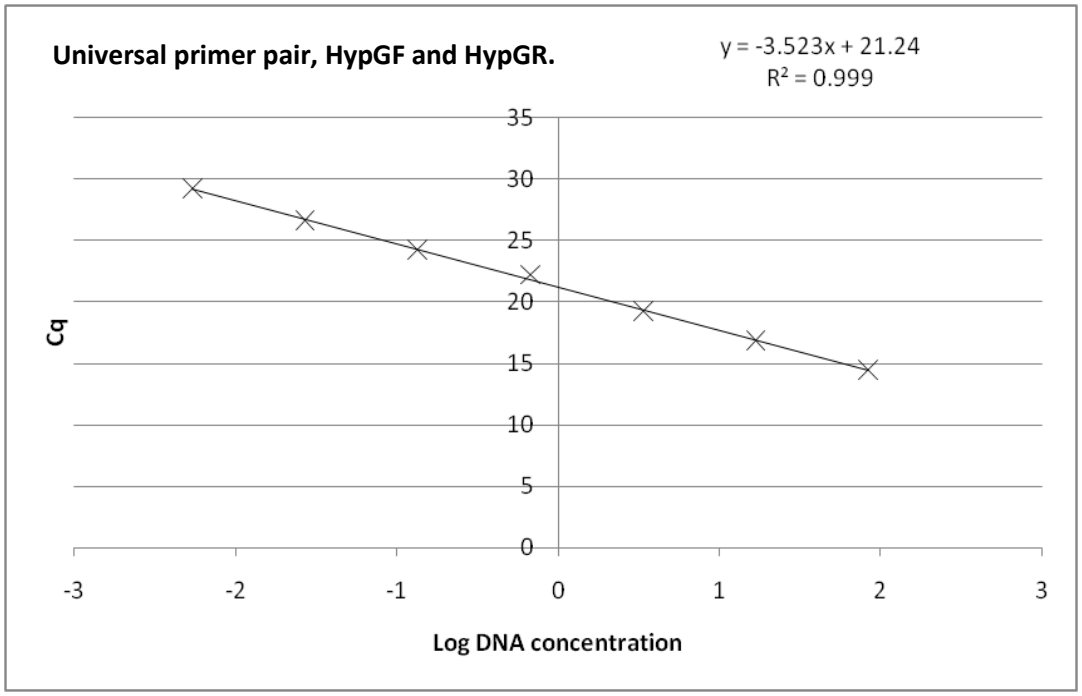


Figure 3.25 Calibration curve for universal primers with *H. perforatum* DNA dilutions, the mean of three Cq values are plotted, each value from a different dilution series listed in Table 3.2.

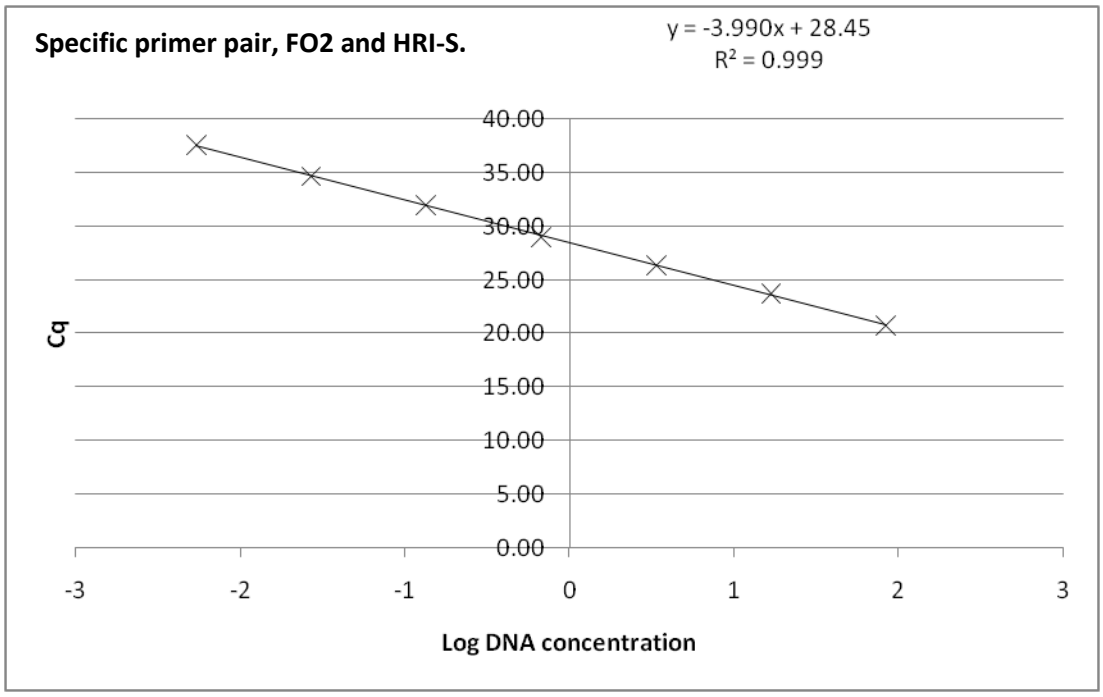


Figure 3.26 Calibration curve of *H. perforatum* specific primers, FO2 and HI-S, with DNA dilutions. The mean of three Cq values are plotted, each value from a different dilution series listed in Table 3.2

3.3.3 *H. perforatum* DNA combined with *H. kouytchense* DNA

An *H. perforatum* DNA dilution series over the same data range was made with *H. kouytchense* DNA (Table 3.3) to simulate the mixed DNA extraction which would be obtained from an adulterated herbal medicinal product. In this series, each data point contained a total of 84ng/µL genomic *Hypericum* DNA, with *H. perforatum* DNA present in the same concentrations as previously and *H. kouytchense* DNA contributing the remainder of the sample. This represents an *H. perforatum* range of 0.01% to 100% of the total DNA content.

The specific primers amplify only the *H. perforatum* DNA, creating a separate calibration curve (Figure 3.27). The values differ slightly from Figure 3.26 due to the presence of the *H. kouytchense* DNA in the reaction, imitating the conditions of medicinal plant adulteration and misidentification. The calibration curve has an R^2 value of 0.995 over a large range of DNA concentrations.

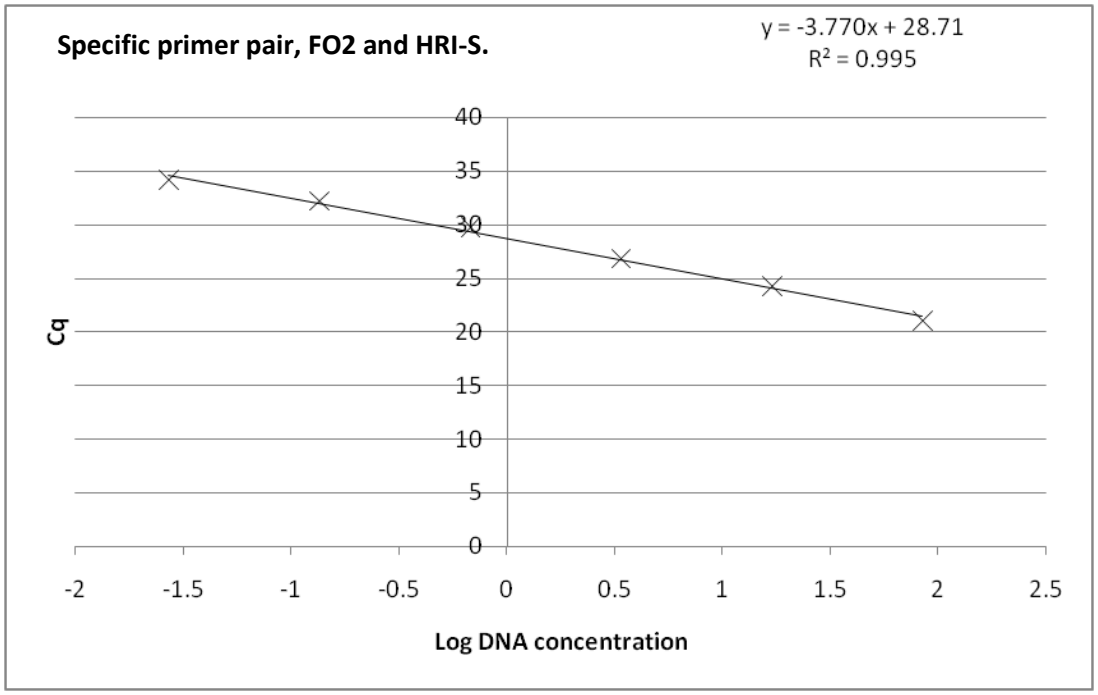


Figure 3.27 Calibration curve for specific primers, FO2 and HRI-S, against *H. perforatum* DNA in a mixed sample. *H. perforatum* and *H. kouytchense* DNA were mixed in a dilution series, (Table 3.3), the specific primers only amplify the *H. perforatum* DNA. The mean Cq values from three replicates with one dilution series are plotted.

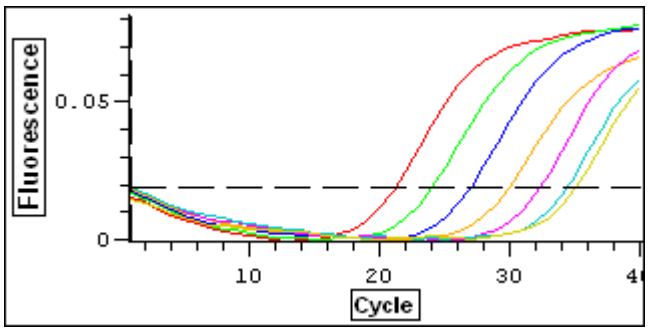


Figure 3.28 One set of qPCR traces from the specific primers, FO2 and HRI-S, with the mixed DNA dilution series PP, described in Table 3.3.

The generic primers HypGF and HypGR amplify all of the DNA present, so each sample has a very similar Cq value ranging from 14.82 to 15.13, with a coefficient of variance of 0.008. The combination of these two assays allows a measure of purity to be calculated for each sample.

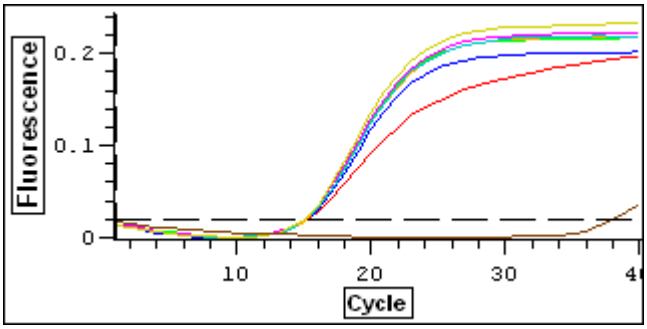


Figure 3.29 qPCR traces from the universal primers, HypGF and HypGR, with the mixed DNA dilution series PP, described in Table 3.3.

Table 3.10 Cq values and range measure for HypGF and HypGR with the PP dilution series described in Table 3.3

Cq Values for all samples
15.13
14.90
15.08
15.08
14.82
15.00
14.84

Mean
14.98
Standard Deviation
0.13
Coefficient of Variance
0.008

3.3.4 Blind Testing

In order to test the theory of measuring purity by dual qPCR assays, a set of samples was produced which contained differing amounts of *H. perforatum* DNA mixed with *H. kouytchense* DNA. The Cq values were then translated into DNA quantities, ng/uL, based on the universal

calibration curve for total DNA content (Figure 3.25) and the specific calibration curve for *H. perforatum* DNA content (Figure 3.27). The figures measured by qPCR closely match the actual DNA concentrations for *H. perforatum* DNA (Table 3.11). However, the total DNA measures are not as close, and two do not follow the same trend as the DNA concentration. This difference results in the % *H. perforatum* DNA in each sample being measured incorrectly, for instance, Sample 1 is 14.75% *H. perforatum*, but is measured as 10.72% *H. perforatum*.

The efficiency values shown are measured based on the slope of the exponential phase of the amplification graph by the Opticon Monitor™ Analysis Software, and show variation between the different samples. This could be part of the cause for the inaccuracies in the qPCR measured results.

Table 3.11 Results from qPCR assay Blind Trials.

Sample	Actual *H. perforatum* DNA, ng/uL	qPCR measure of *H. perforatum* DNA, ng/uL	qPCR Efficiency, %	Total DNA, ng/uL	qPCR measure of total DNA ng/uL	qPCR Efficiency, %
1	3.07	2.68	50.46	20.82	25.00	82.68
2	0.38	0.39	50.66	18.13	24.35	97.54
3	0.048	0.05	50.79	17.80	27.75	83.49
4	0.006	0.03	55.00	17.76	27.57	87.30

3.3.5 Discussion

The qPCR method is capable of accurately measuring both total DNA and specific *H. perforatum* DNA within the same sample, when those DNA samples have been used to construct a calibration curve. The inability to transfer qPCR calibrations between different laboratories has been addressed with the use of synthetic external oligonucleotide standards (Vermeulen et al., 2009). However, different DNA extractions alter results in unexpected ways which in this situation prevents this method being applicable to industry. Two of the main reasons for this are the difficulty of meaningful DNA quantification and the measurement of and allowance for differences in qPCR efficiency.

3.3.5.1 DNA Quantification

In order to conduct a qPCR experiment a reliable and reproducible method of measuring DNA concentration is required to enable production of the calibration curves, and with which to compare qPCR results. The most commonly used method is measuring absorbance at 260nm, which can then be compared with the absorbance at 280nm to give a measure of purity.

However, this method measures total nucleotide content of a sample which may include single-stranded DNA, RNA and DNA degradation products.

The use of fluorophores to measure DNA confers specificity to double-stranded DNA (Nielsen et al., 2008). The QuantIT™ (full details in section 3.2.1) system used in this investigation contains the DNA intercalating dye PicoGreen® and is an example of this type of assay. However, the DNA concentration measured is dependent on factors including temperature and incubation time, and can vary within a single sample if three readings are taken over a period of approximately one minute. This technique will also include any sheared double-stranded DNA, and DNA from plastid and mitochondrial genomes which is not then measured in the qPCR amplification. This would not be of concern if it could be guaranteed that all samples contain the same ratio of plastid and mitochondrial DNA to genomic DNA.

It could be argued that the qPCR itself is the most meaningful measure of the DNA concentration, as it is specific to amplifiable quality double-stranded DNA. However, the use of the same method to both calibrate and test an assay is bad practice, and could very quickly introduce systematic errors.

One possible method to overcome this is to standardise DNA samples prior to qPCR by amplifying the region of interest and introducing only this to qPCR. This would eliminate any single-stranded or degraded DNA and remove DNA from other genomes. However, whether this would still give a representative picture of the original DNA sample, and therefore the sample it was extracted from, would require very careful investigation.

3.3.5.2 qPCR Efficiency

qPCR efficiency is described as the rate of amplification, and can be given as a percentage value or a value with a maximum of 1 depending on preference. The theoretical maximum of either 100% or 1 is achieved when the amount of DNA doubles with each cycle, this occurs in the exponential phase of the reaction so efficiency can be measured by the slope of the fluorescence curve in this phase. The MIQE preferred method of measuring efficiency is based on the slope of the calibration curve, where Cq is plotted against the log of the DNA concentration (Bustin et al., 2009).

Table 3.12 qPCR Efficiencies of calibrations curves as measured by the slope of the curve for universal and specific primers.

Calibration Curve	Primer Pair	Slope	% Efficiency
Figure 3.25	Universal	-3.5230	92.24
Figure 3.26	Specific	-3.9900	78.08
Figure 3.27	Specific	-3.7700	84.18

The efficiency of a qPCR reaction is a function of the primers used, both in terms of their annealing and the template DNA sequence which will be present in between them after amplification. A longer amplicon is more time consuming to replicate than a shorter one, so an optimal amplicon length is essential in qPCR to ensure that entire copies can be produced in each cycle. The size of the amplicon affects the efficiency regardless of the method of measurement, but it also alters the Cq value. The copying of a shorter amplicon will cause a smaller increase in fluorescence than the copying of a longer amplicon, altering the number of cycles necessary to cross the threshold.

qPCR efficiency alters with different DNA extractions (Table 3.12 and Table 3.11), requiring different calibration curves, though this can be overcome by the use of reference genes. The dual amplification of two genes is used to standardise any differences between DNA extractions. The effect of any difference is considered to be consistent between both the gene of interest and the reference gene. Any deviation in that difference is then due to the concentration of the target gene altering in relation to the reference, this is termed the ΔΔCq (Livak and Schmittgen, 2001, Bustin et al., 2009), the change in the difference between the target and reference gene Cq. However, the ΔΔCq method for quantifying difference requires that the two genes are amplified with identical efficiency, which is not the case when comparing the universal and specific primers in this chapter. The MIQE (Bustin et al., 2009) recommended formula for use when comparing target and reference amplicons with different qPCR efficiencies was created by Pfaffl;

$$\textbf{Ratio} = \frac{(E_{\text{target}})\,\Delta Cq_{\text{target}}\ \textit{(control-sample)}}{(E_{\text{ref}})\,\Delta Cq_{\text{ref}}\textit{(control-sample)}}$$

(Pfaffl, 2001)

Where ‘E target’ is the efficiency of the target gene or region, ‘E ref’ is the reference gene or region and the Cq of each is compared to a control. In this example, the target region is the *H. perforatum* specific amplicon, the reference is the universal amplicon. Separately for the target and the reference, the controls are one of the calibration curve measurements and the samples are to be measured. However, this assumes that the efficiencies of the controls are equal to those of the samples, which has not been the case in the assay described in this chapter.

3.3.5.3 Conclusion

The aims of this research were two fold;

- That *H. perforatum* DNA could be accurately measured by qPCR to enable a quantitative measure
- That a dual assay might be capable of measuring the purity of an *H. perforatum* sample, in line with the requirements of the European Pharmacopoeia

The calibrations for both the specific and the universal qPCR reactions showed very high correlation to the measured DNA, proving that the technique can be applied to medicinal plant DNA. The specific qPCR reaction showed very close results to the measured DNA content of the samples. This showed that *H. perforatum* could be accurately detected and measured within a mixed sample. This also furthers the possibilities of application of the microcode PCR assay, as qPCR is a quicker process than conventional PCR as it does not require the second stage of gel electrophoresis. The use of analysis software also enables qPCR to become a high throughput technique, allowing 96 samples to be measured simultaneously on an average machine.

The universal qPCR measured the calibration samples very accurately, but was more susceptible to the change in DNA extraction than the specific qPCR reaction. The concentrations of DNA measured for each sample in the blind trials were similar to the actual measures, though not as accurate as previous samples and two did not follow the correct trend of increasing concentration.

The application of this dual qPCR assay to mixed DNA samples containing *H. perforatum* and another *Hypericum* DNA is capable of detecting contamination. However, the European Pharmacopoeia requires that SJW medicinal products be 98% pure. As this qPCR method stands, it is not capable of measuring to this degree.

This may be due in part to the methodological design of the assay as one of the greatest strengths of qPCR is the extremely low theoretical level of detection; due to this it is much more powerful when used to measure and detect very low levels of DNA. For example, the specific qPCR reaction for *H. perforatum* measured accurately down to 0.05ng/uL (Table 3.11).

The scale of the Cq measure is not linear, a difference in Cq of 0.1 at cycle number 10 has a much greater effect on the subsequent quantification than the same difference at cycle number 35. A more useful design for this technique would therefore be to detect very low levels of adulterant or dangerous material.

The dual qPCR assay described in this chapter was also constrained by the features and variations of the nrITS region, used as a platform for design. It may be possible in future to use different regions, which may contain sequences which enable the design of exactly matched primers and amplicons to overcome the efficiency problems discussed.

4 Identification of Individual Species within a Mixed Sample

4.1 Introduction

4.1.1 Multiple species in one sample

There are many reasons for the presence of plant material from many different species being present in one sample. Generally, these fall into three categories:

- Misidentification, leading to accidental substitution of plant material
- Intentional substitution or adulteration, for economic or availability reasons
- Intentional blending of plants for the benefit of the user

4.1.1.1 Misidentification and substitution

The dangers of the first two situations are that the plant used in the place of the correct species may well not have the medicinal activity the user intends. This may mean that there is no biological activity, or, more worryingly, that the plant has detrimental health implications.

Situations in which a dangerous substance is used in the place of a medicinal plant can put the safety of the user at risk. This has been described for the cases of *Aristolochia* and *Illicium* in sections 1.1 and 1.3.3 respectively. Another noteworthy potential example of this is Black Cohosh, *Actaea racemosa*, used to reduce menopausal symptoms. The use of this plant has been linked with cases hepatotoxicity, but products labelled as "Black Cohosh" have been found not to contain any of the correct plant material. This has led to the suspicion that the cases of adverse reactions may in fact be linked to cases of substitution or adulteration, rather than to *A. racemosa* (Jordan et al., 2010).

Particularly when misidentification and adulterations of this nature are known to occur, identification methods must not only verify the presence of the intended plant material, but also eliminate the possibility of adulteration with a harmful substance.

4.1.1.2 Synergy and Polyherbal Preparations

Synergy can be simply described as cooperation between two entities resulting in an enhanced outcome, the whole being greater than the sum of its parts.

In section 1.4.2, the bio-active compounds of SJW were discussed, leading to the finding that only the whole extract can be considered the active compound. This is an example of synergy within one medicinal plant, where the compounds found within one plant extract or preparation operate with a concerted action (Williamson, 2001). Also common within

phytomedicines is the use of combinations of medicinal plant material, which is fundamental to the practices of both TCM and Ayurveda. Novel mixtures with specific applications are now patented, e.g. BHUx a blend of five plants from Ayurveda, used for the prevention of atherosclerosis (Tripathi et al., 2005) and Glyoherb a polyherbal combination for the treatment of diabetes (Thakkar and Patel, 2010).

The new European Legislation (Directive 2004/24/EC) requires that where multiple components are included in a preparation, each must be shown to be efficacious (Wagner, 2009). Extending from these situations, it is likely that a preparation will apply for Registration containing many medicinal plant constituents, one or many of which must be considered as a whole plant, as SJW is.

4.1.2 Aims and Objectives

Dangerous adulterants and synergistic polyherbal formulas present similar identification and authentication requirements. Multiple plant species must be identified, whether this is because their presence is harmful or beneficial to the preparation.

A DNA based method for the detection of multiple mammal species was developed by Tobe and Linacre (Tobe and Linacre, 2008) in which species specific PCR reactions for 18 mammals are conducted in one reaction, a multiplex reaction. Using the cytochrome *b* gene as a design platform, three universal forward primers were produced and 35 species specific reverse, each combination resulting in an amplicon of a different size. The three universal forward primers were labelled with fluorescent dyes, enabling detection of the amplicons by capillary electrophoresis. The result from this is a peak pattern on an electropherogram in which each peak indicates the presence of the DNA of a different species.

The aim of the current work was to develop a similar method to discriminate different *Hypericum* species, using the nrITS region sequences. As in Chapter 2, these regions were used as a model for barcoding data which could become a platform for the design of these methods for any and all medicinally important and economically valuable plants, and to aid in the difficult circumstances described above.

4.2 Materials and Methods

4.2.1 Multiplex PCR Primer Design

Primers were designed by Dr Adrian Slater using Allele ID software (Apte and Singh, 2007) available at http://www.premierbiosoft.com/bacterial-identification/index.html, input sequence GenBank Accession numbers shown in Table 4.1.

Table 4.1 GenBank accession numbers and species names for sequences used in the design of pimers with AlleleID

Accession number	Species
EU796888	*H. perforatum*
AF455674	*H. perforatum*
AJ414728.1	*H. calycinum*
AY555839.1	*H. perforatum*
AY555842.1	*H. maculatum*
AY555843.1	*H. graveolens*
AY555844.1	*H. punctatum*
AY555845.1	*H. delphicum*
AY555846.1	*H. athoum*
AY555849.1	*H. ascyron*
AY555853.1	*H. kouytchense*
AY555861.1	*H. calycinum*
AY572993.1	*H. attenuatum*
AY573012.1	*H. androsaemum*
AY573014.1	*H. ascyron.*

4.2.2 DNA Template Preparation

DNA samples used as templates for initial primer testing were supplied by the Royal Botanic Gardens Kew, as listed in Section 2.2.2. The nrITS regions were amplified using primers and cycling conditions listed in Section 2.2.4. PCR reactions were conducted with GeneAmp® High Fidelity PCR System (Applied Biosystems, Foster City, CA), final volume 50µL. Reactions, in 0.2mL polypropylene tubes consisted of GeneAmp High Fidelity PCR Buffer (without $MgCl_2$) (1X), $MgCl_2$ (2.5mM), GeneAmp High Fidelity Enzyme Mix (2.5 Units), relevant primers (0.1µM

each), dNTPs (0.1µM each), nuclease-free water and template DNA (0.7-1µg). Amplifications were carried out personally and by a Research Assistant (Sarah Smith) under supervision.

Products from this amplification were diluted as shown in Table 4.2 to provide target species template for primer validation.

Table 4.2 Initial PCR dilutions for primer testing against target species

Species	Initial PCR number	Dilution 1	Dilution 2	Label
H. androsaemum	1	10^{-5}	1 in 6	and1.5.6
H. kouytchense	2	10^{-5}	1 in 6	kou2.5.6
H. maculatum	4	10^{-5}	1 in 6	mac4.5.6
H. athoum	8	10^{-4}	1 in 6	ath8.4.6
H. calycinum	9	10^{-5}	1 in 6	cal9.5.6
H. ascyron	12	10^{-5}	1 in 6	asc12.5.6
H. perforatum	3, 7, 10	10^{-5}	1 in 6	perfmix.5.6

These amplifications were also used to prepare "non-target DNA panels". 10µL of dilution 1 for each required species was added as described in Table 4.3. Each individual DNA within the panel was then at the same concentration as each target DNA sample.

Table 4.3 Panel construction for non-target DNAs.

Panels are described by the name of the species which is not present, column one. The different DNA amplifications used to make up the panels are shown, the names refering to dilution 1 in Table 4.2.

Description	and 1.4	kou 2.4	mac 4.4	ath 8.3	cal 9.4	asc 12.4	perf mix.4	Label
Non-and	x	✓	✓	✓	✓	✓	✓	$P1^{-1}$
Non-kou	✓	x	✓	✓	✓	✓	✓	$P2^{-1}$
Non-mac	✓	✓	x	✓	✓	✓	✓	$P3^{-1}$
Non-ath	✓	✓	✓	x	✓	✓	✓	$P4^{-1}$
Non-cal	✓	✓	✓	✓	x	✓	✓	$P5^{-1}$
Non-asc	✓	✓	✓	✓	✓	x	✓	$P6^{-1}$
Non-perf	✓	✓	✓	✓	✓	✓	x	$P7^{-1}$

4.2.3 Initial Primer Testing

Primers designed were as listed in Appendix Section 8.1.2; each recommended combination (Table 4.4) was first tested against target DNA nrITS amplifications. Reactions were conducted personally and by a Research Assistant (Sarah Smith) under supervision.

PCR reactions consisted of Green GoTaq® Flexi Buffer (Promega, Madison, WI, USA) (1x), $MgCl_2$ (2.5mM), GoTaq® DNA Polymerase (Promega) (1.25 Units), relevant primers (0.1µM each), dNTPs (0.1µM each), and template DNA (1µL of appropriate sample dilution, listed in Table 4.2) made up to a final volume 25µL with nuclease-free water in 0.2mL polypropylene tubes. The Applied Biosystems GeneAmp PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA) was used with the programme: 7min at 95°C initial denaturation step, 30 cycles consisting of 1min at 95°C, 30s at 60°C and 1min at 72°C, final extension period of 7min at 72°C.

Combinations requiring further optimisation were run on a gradient of annealing temperatures from 55°C to 69°C, with all other parameters as described above.

Reactions without template DNA were utilised as controls. PCR products were run on 50mL 3% (w/v) agarose, 0.5 X TBE gels with 1µL SYBRsafe™ (Invitrogen, Carlsbad, CA, USA) DNA stain at 90V for ~30min and analysed in a BioRad Illuminator with ChemiDocXRS Camera and Quantity One software.

Table 4.4 Recommended primer combinations with individual *Tm*s and product lengths

Forward Primer name				Tm °C	Reverse Primer name				Tm °C	Product Length bp
Hand	F	1	1	67.0	Hand	R	1	1	67.4	104
Hand	F	1	2	67.0	Hand	R	1	2	66.5	65
Hand	F	1	3	64.7	Hand	R	1	3	64.8	135
Hand	F	1	4	65.6	Hand	R	1	4	65.9	67
Hand	F	1	4	65.6	Hand	R	1	5	65.4	63
Hand	F	1	5	67.1	Hand	R	1	6	67.3	98
Hand	F	2	1	65.1	Hand	R	1	3	64.8	137
Hasc	F	1	1	60.9	Hasc	R	1	1	60.4	221
Hasc	F	2	1	57.5	Hasc	R	2	1	57.5	213
Hasc	F	1	2	60.9	Hasc	R	1	2	61.7	225
Hasc	F	1	4	62.5	Hasc	R	1	2	61.7	231
Hasc	F	1	4	62.5	Hasc	R	1	1	60.4	224
Hasc	F	1	3	61.5	Hasc	R	1	1	60.4	219
Hasc	F	1	5	60.2	Hasc	R	1	1	60.4	217
Hasc	F	2	2	61.0	Hasc	R	2	2	60.9	228
Hasc	F	2	2	61.0	Hasc	R	2	3	62.1	234
Hasc	F	2	3	56.2	Hasc	R	2	4	57.7	222
Hasc	F	2	4	59.6	Hasc	R	2	5	59.3	73
Hasc	F	1	1	60.9	Hasc	R	2	2	60.9	238
Hath	F	1	1	61.8	Hath	R	1	1	61.7	137
Hath	F	1	1	61.8	Hath	R	1	3	64.6	133
Hath	F	1	2	69.7	Hath	R	1	2	70.7	148
Hath	F	1	1	61.8	Hath	R	1	4	64.9	127
Hath	F	1	3	63.2	Hath	R	1	3	64.6	151
Hath	F	1	4	65.8	Hath	R	1	4	64.9	146
Hcal	F	2	1	59.7	Hcal	R	2	1	59.6	240
Hcal	F	2	2	61.0	Hcal	R	2	1	59.6	238
Hcal	F	2	3	59.6	Hcal	R	2	2	59.7	128
Hcal	F	2	4	59.2	Hcal	R	2	1	59.6	224
Hcal	F	2	4	59.2	Hcal	R	2	3	59.4	225
Hcal	F	2	1	59.7	Hcal	R	2	3	59.4	241
Hkou	F	1	1	79.8	Hkou	R	1	1	65.8	179
Hkou	F	1	1	79.8	Hkou	R	1	2	64.9	175
Hkou	F	1	1	79.8	Hkou	R	1	3	65.7	182
Hmac	F	2	1	65.1	Hmac	R	2	1	64.4	240
Hmac	F	2	1	65.1	Hmac	R	2	2	65.5	231
Hmac	F	2	1	65.1	Hmac	R	2	3	64.9	236
Hmac	F	2	2	66.4	Hmac	R	2	2	65.5	235
Hmac	F	2	3	66.9	Hmac	R	2	1	64.4	242
Hmac	F	2	3	66.9	Hmac	R	2	2	65.5	233
Hper	F	1	1	69.0	Hper	R	1	1	65.0	273
Hper	F	3	1	67.3	Hper	R	3	1	66.7	60
Hper	F	4	1	65.1	Hper	R	4	1	64.9	222
Hper	F	1	2	72.5	Hper	R	1	1	65.0	281
Hper	F	1	3	71.9	Hper	R	1	1	65.0	277
Hper	F	1	4	71.2	Hper	R	1	1	65.0	275

4.2.4 Multiplex PCR and Capillary Electrophoresis

Reactions and analysis were conducted personally and by Eleni Socratous and Dr Eleanor Graham of the East Midlands Forensic Pathology Unit, Leicester University.

Potential interactions between primers were assessed using AutoDimer v.1 Software (Vallone, 2004) available at http://yellow.nist.gov:8444/dnaAnalysis/primerToolsPage.do.

Multiplex PCR reactions, final volume 10µL, were conducted in 0.2mL polypropylene tubes using the Qiagen Multiplex PCR Kit (Qiagen Inc., CA). Reactions consisted of Multiplex PCR Master Mix (1X) (Qiagen), primer mix (varied between 80 and 500nM each) (VHBio, Gateshead UK or IDT, Iowa USA) template PCR product (0.4µL at working concentration) and nuclease-free water. (Fluorescently labelled primers were from Invitrogen and are listed in Table 4.7)After optimisation, the concentration of each primer in a pair was; *H. perforatum* – 500nM, *H. athoum* – 110nM, *H. androsaemum* – 80nM, *H. ascyron* – 400nM. Reactions without template PCR product were used as controls. Cycling parameters were: Initial denaturation step at 95^{o}C for 15min; 30 cycles of 94^{o}C for 30s, 64^{o}C for 90s, 72^{o}C for 60s; final extension 30min at 60^{o}C.

Products were analysed at the East Midlands Forensic Pathology Unit on the ABI Prism™ 3130 Genetic Analyzer (Applied Biosystems, Foster City, CA), using a 30cm capillary and Performance Optimised Polymer 4 (Applied Biosystems, Foster City, CA). The run module used consisted of a 12s injection at 1.2kV, followed by electrophoresis running at 60^{o}C and 15kV for 25min. 1µL of multiplex PCR product was diluted with 8.5µL Hi Di™ Formamide and 0.5µL GeneScan™ -500 ROX™ size standard (Applied Biosystems, Foster City, CA) before capillary electrophoresis. GeneMapper® ID v3.2 fragment analysis software (Applied Biosystems, Foster City, CA) was used, all settings were default.

4.3 Results and Discussion

4.3.1 Primer Design

The primers used in this system were designed using AlleleID software (See section 4.2.1 for details). This software aligns input sequences and analyses them to locate areas of sequence difference for each of the input sequences as compared to the others. PCR primers are then designed in these areas; the output from this is a list of proposed pairings which are rated as to their quality, only pairs rated as 'good' or 'best' were investigated further, with the exception of the *H. kouytchense* primers as so few were designed that 'poor' pairs had to be tested.

The input sequences had to be grouped in order for primers to be designed for each species, as the AlleleID program could not recommend primers for every species in the complete alignment of sequences from 8 species. The species for which primers were designed from the complete alignment were placed in Group 1. Species in this group were then removed sequentially from the alignment until the program was able to design primers to the all of the species not placed in Group 1. Group 3 was a selected sub-group designed to distinguish the "Crockett" *Hypericum perforatum* subsp. *perforatum* isolate hyp1ITS from its putative parent *H. attenuatum, and Group perf was designed to distinguish the different Hypericum perforatum* ITS sequences found in the GenBank database. The group is referenced in the name of each primer, i.e. the primer named Hper F.1.2 is a primer for *H. perforatum*, in the forward direction, designed against sequences in Group 1, and it is the second primer designed. All other sequences in that group are termed 'anti-targets', as the primers are designed particularly not to amplify those sequences.

Table 4.5 Groups of DNA sequences input into AlleleID software in order to design primers for the PlantID system.

Group 1
Hypericum perforatum
Hypericum attenuatum
Hypericum athoum isolate hyp8ITS
Hypericum ascyron isolate hyp11ITS
Hypericum kouytchense isolate hyp15ITS
Hypericum androsaemum
Group 2
Hypericum ascyron isolate hyp11ITS
Hypericum calycinum isolate hyp23ITS
Hypericum androsaemum
Hypericum maculatum isolate hyp4ITS
Group 3
Hypericum perforatum subsp. *perforatum* isolate hyp1ITS
Hypericum attenuatum
Group perf
Hypericum perforatum subsp. *perforatum* isolate hyp1ITS
Hypericum perforatum

4.3.2 Template DNA for Initial Primer Testing

Obtaining sample DNA for each of the species that the assay was designed for proved to be problematic. The vouchered DNA samples obtained from the Kew DNA Bank were strictly limited. Seeds were available to purchase from on-line retailers, but delays in delivery and germination caused significant time delays. In order to circumvent this problem, amplified nrITS regions were produced as templates from the vouchered DNA, enabling empirical testing to be conducted, while reserving limited stocks of the remaining vouchered DNA samples.

PCR reactions are typically performed with *Taq* polymerase which has a low but significant error rate, as it does not have a 3′ to 5′ proofreading activity. *Taq* polymerase inserts incorrect bases during elongation at a rate of 1×10^{-5} errors per base, whichs is not a general concern as the end product is not usually used for anything other than gel electrophoresis analysis. As the PCR products in this case were used to verify primers in further PCR reactions, an error

introduced into the sequence of the amplicon could cause misleading results, particularly if this error occurred early on in the reaction and became the template for a significant number of amplicons. This situation was avoided by the use of a High Fidelity PCR product which does have a proofreading enzyme. This exonuclease activity reads the DNA products in a 3′ to 5′ direction and corrects any bases which have been incorporated incorrectly, conferring an accuracy six times higher than *Taq* polymerase (Promega, 2009).

The initial nrITS amplification products, Figure 4.1, were then diluted to a concentration equivalent to genomic DNA. The working concentration of these PCR products required for further testing was determined using several dilutions of the *H. perforatum* 13932 and the *H. androsaemum* amplifications, and the *H. perforatum* specific primers FO2 and HRI-S. The kinetics of PCR reactions are greatly affected by template DNA concentration, as the reaction can be saturated, forcing amplification when overwhelming concentrations of template DNA are present. This can be seen in the most concentrated DNA samples in Figure 4.2, panel A, as *H. androsaemum* DNA acts as a template for *H. perforatum* specific primers. Figure 4.2 shows that the optimal dilution a target DNA is 10^{-5} as this is the lowest concentration at which the specific product is formed, while cross amplification ceases to occur at 10^{-3}.

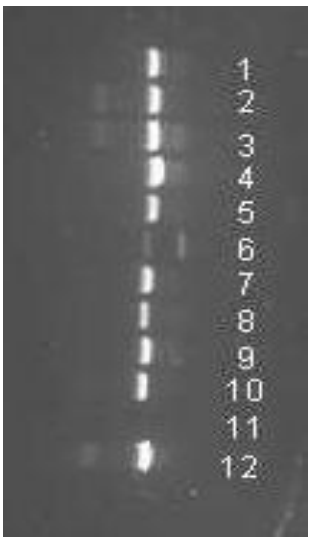


Figure 4.1 PCR products from nrITS amplification.

Samples are as follows: 1 *H. androsaemum*, 2 *H. kouytchense*, 3 *H. perforatum* 13876, 4 *H. maculatum*, 5 *H. patulum*, 6 Hidcote, 7 *H. perforatum* 13921, 8 *H. athoum*, 9 *H. calycinum*, 10 *H. perforatum* 13932, 11 *H. delphicum*, 12 *H. ascyron*.

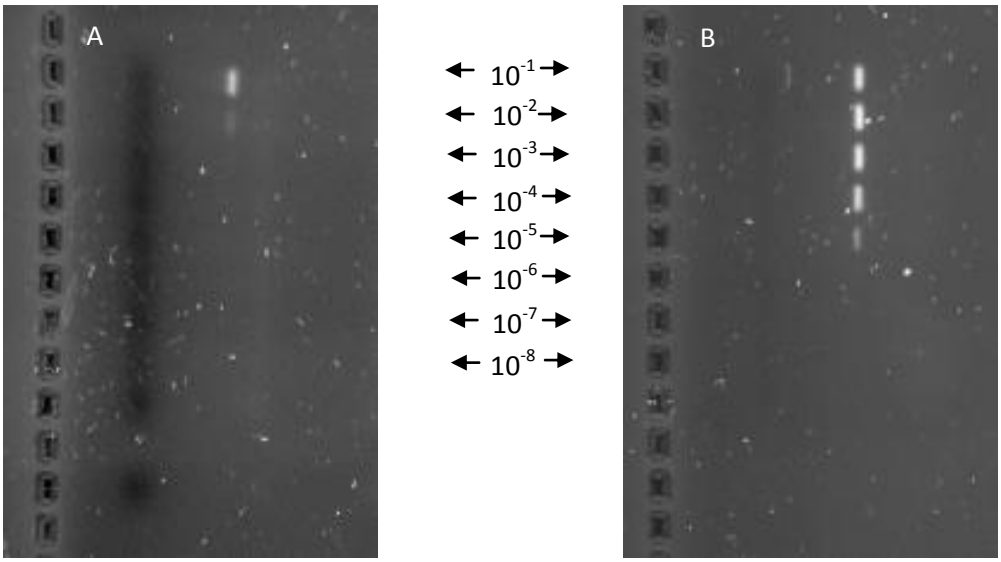


Figure 4.2 PCR products from FO2 and HRI-S with dilutions of nrITS amplifications.

Dilution factors indicated are relevant to both gel images. Panel A shows results using dilutions of *H. androsaemum* PCR products and Panel B shows *H. perforatum* 13932 PCR products. Products are no longer formed at dilution factor 10^{-3} in panel A but continue to be produced up to 10^{-5} in panel B.

4.3.3 Initial Primer Testing and Selection

Initial primer testing was carried out with two objectives;

- To ensure a positive reaction with target DNA, i.e. the target DNA that the primers were designed to amplify
- To ensure a negative reaction with non-target DNA, i.e. no product with DNA from any of the other species in the design, the non-target DNA panel

Cross-amplification was assessed by PCR reactions with DNA panels made up of non-target DNA (Table 4.3), each contained DNA from six species. Three accessions of *H. perforatum* DNA were used to allow for any variations within this target species, which have been shown to occur (Section 5.3.1). As the DNA samples in the panels each accounted for one sixth of the total DNA present, the template used for target DNA was further diluted to match the concentration of each individual in the panel (Table 4.2).

Each recommended primer pair was then tested against its target DNA, and the corresponding non-target panel. A selection of the results is shown in Figure 4.3; each primer pair tested in this example produced a single amplicon of the expected size, showing the efficiency of the AlleleID software. Figure 4.4 shows the results from a gradient run with three different primer pairs. In the qPCR chapter, gradient runs were used to identify the optimal annealing temperatures for primer pairs (Section 3.3.1). Here gradient runs are used to increase the stringency of the PCR reaction to a level where only the target amplicon is produced and no cross amplification occurs. In Figure 4.4, this separation only transpires for one primer pair, indicated by arrows on the Figure.

The process of testing and selection was carried out for all 46 recommended primer combinations and their respective target DNA and non-target panel DNA. The full results are listed the Appendix, Section 8.1.3, and the resultant candidate primers are shown in Table 4.6.

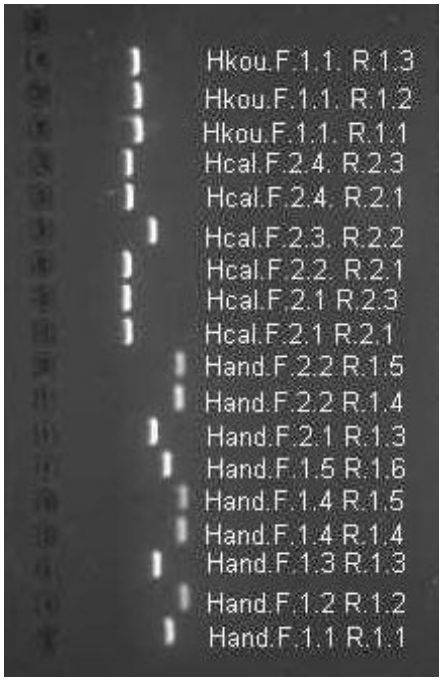


Figure 4.3 PCR products from a selection of the primer pairs.

Primers are as shown, and each reaction was with the relevant target DNA, as indicated by the names of the primers. All primer pairs produced the intended target product, as designed by the AlleleID software.

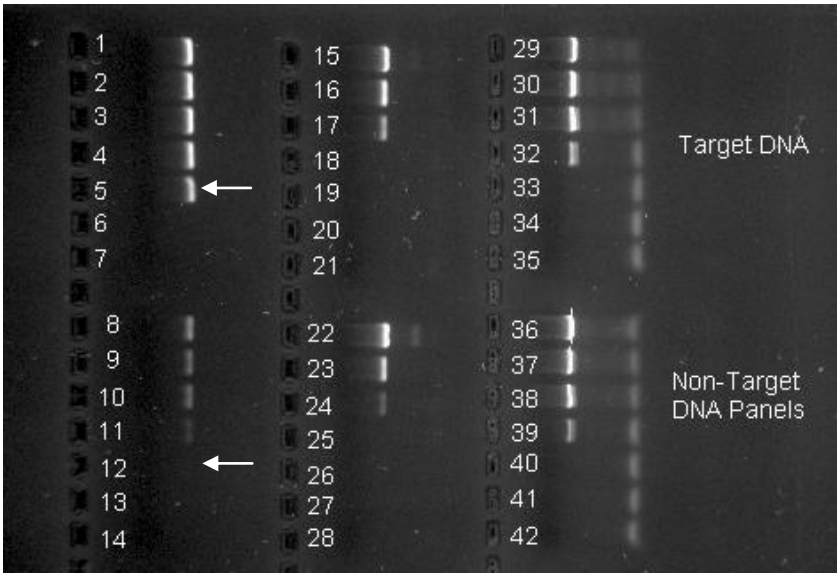


Figure 4.4 Gel image of PCR products from three primer pairs.

Lanes 1-7 are primers HandF.1.1 and HandR.1.1 with target DNA, 8-14 with the Non-androsaemum Panel; Lanes 15-21 are primers HcalF.2.3 and Hcal R.2.2 with target DNA, 22-28 with the Non-calycinum Panel; Lanes 29-35 are primers HkouF.1.1 and HkouR.1.3 with the target DNA, 36-42 with the Non-target Panel. For each set of seven reactions, all parameters were equal other than the annealing temperature, which ranged as follows: 1 - 60°C, 2 - 62°C, 3 – 63.4°C, 4 – 65.2°C, 5 – 67.2°C, 6 – 70.2°C and 7 - 72°C. In all cases, as the annealing temperature increases the amount of product reduces to the point where they are not visible. To be a candidate primer pair, the product with the non-target panel DNA must be prevented from forming while the product with the target DNA remains. This happens with one primer pair, HandF.1.1 and HandR.1.1 at 67.2°C, as indicated by arrows.

Table 4.6 Candidate primer pairs after testing with target DNA and non-target panels.

Candidates highlighted are of a beneficial length for separation of products via capillary electrophoresis. Candidates marked Y give a target product and do not cross-amplify, Y – in theory indicates that previous results have shown the pair to be specific at lower stringency so can be assumed to be specific at the given parameters.

Forward Primer name				*Tm*	Reverse Primer name				*Tm*	Product Length	Candidate @ 64°C?	Candidate @ 63°C?
Hand	F	1	2	67.0	Hand	R	1	2	66.5	65	Y	Y
Hand	F	1	3	64.7	Hand	R	1	3	64.8	135	Y	Y
Hand	F	1	4	65.6	Hand	R	1	4	65.9	67	Y	Y
Hand	F	1	4	65.6	Hand	R	1	5	65.4	63	Y	Y
Hand	F	1	5	67.1	Hand	R	1	6	67.3	98	Y	Y
Hand	F	2	1	65.1	Hand	R	1	3	64.8	137	Y	Y
Hasc	F	1	2	60.9	Hasc	R	1	2	61.7	225	N	Y
Hasc	F	1	4	62.5	Hasc	R	1	2	61.7	231	Y	Y
Hasc	F	1	3	61.5	Hasc	R	1	1	60.4	219	N	Y
Hasc	F	1	5	60.2	Hasc	R	1	1	60.4	217	N	Y
Hath	F	1	1	61.8	Hath	R	1	1	61.7	137	Y	Y- in theory
Hath	F	1	1	61.8	Hath	R	1	3	64.6	133	Y	Y- in theory
Hath	F	1	1	61.8	Hath	R	1	4	64.9	127	Y	Y- in theory
Hath	F	1	3	63.2	Hath	R	1	3	64.6	151	N	Y
Hper	F	1	1	69.0	Hper	R	1	1	65.0	273	Y	Y- in theory
Hper	F	4	1	65.1	Hper	R	4	1	64.9	222	Y	Y- in theory
Hper	F	1	2	72.5	Hper	R	1	1	65.0	281	Y	Y- in theory
Hper	F	1	3	71.9	Hper	R	1	1	65.0	277	Y	Y- in theory
Hper	F	1	4	71.2	Hper	R	1	1	65.0	275	Y	Y- in theory

Candidate primer pairs which fulfilled the requirements of target amplification and no cross-reactions were found for four of the original seven species included in the design, the three species without candidate primer pairs are; *H. kouytchense*, *H. maculatum* and *H. calycinum*.

In the case of *H. maculatum*, the annealing temperature required to prevent cross-amplification also prevented the formation of the target amplicon in all primer combinations. The *H. calycinum* primers cross-reacted with one or more of the DNA templates in the Non-cal panel, which may be expected as no primers were generated for this species when all of the Group 1 sequences were input into the software due to insufficient sequence differences.

One forward primer was generated to target the *H. kouytchense* sequence, and this had a *Tm* of 79.8^{o}C (Table 4.4). Increasing the annealing temperature to 67^{o}C did not prevent cross-amplification. As shown in Figure 4.5, products are formed with the non-target panel of DNA. As the final technique required a multiplex PCR reaction, the testing annealing temperature could not reasonably be increased any further as this would be likely to prevent the other primers in the reaction from annealing to their targets.

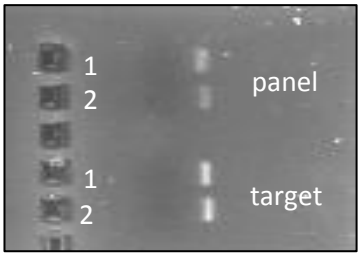


Figure 4.5 Image of gel with amplicon bands produced with *H. kouytchense* primers.
HkouF.1.1 and HkouR.1.2 in lane 1 and HkouF.1.1 and HkouR.1.3 in lane two. Above; the template DNA was the non-target panel N-kou, Below; the template DNA was the target kou2.5.6. A band is produced in all combinations, so the primer pairs must be discounted.

All of the candidate primers for four species are shown in Table 4.6, those highlighted produce amplicons of different sizes making them beneficial for separation after the multiplex reaction.

4.3.4 Multiplex PCR

Of the candidate primer pairs found for each of the four species, one pair per species was selected for the multiplex reaction (Table 4.7). These were chosen based on the analysis of AutoDimer v.1 software (Full details in Section 4.2.4), which highlighted the possibility of interactions between candidate primers when all were introduced into the multiplex system.

Figure 4.6 shows an example of the AutoDimer analysis - primer interactions are scored allowing easy selection of the most suitable primers to be used.

Table 4.7 Primers selected to be fluorescently labelled and used in the Multiplex reaction to be detected by capillary electrophoresis.

Forward Primer	Reverse Primer	Amplicon Length, bp
Hper.F.4.1	Hper.R.4.1	222
Hand.F.1.4	Hand.R.1.4	67
Hath.F.1.3	Hath.R.1.3	151
Hasc.F.1.4	Hasc.R.1.2	231

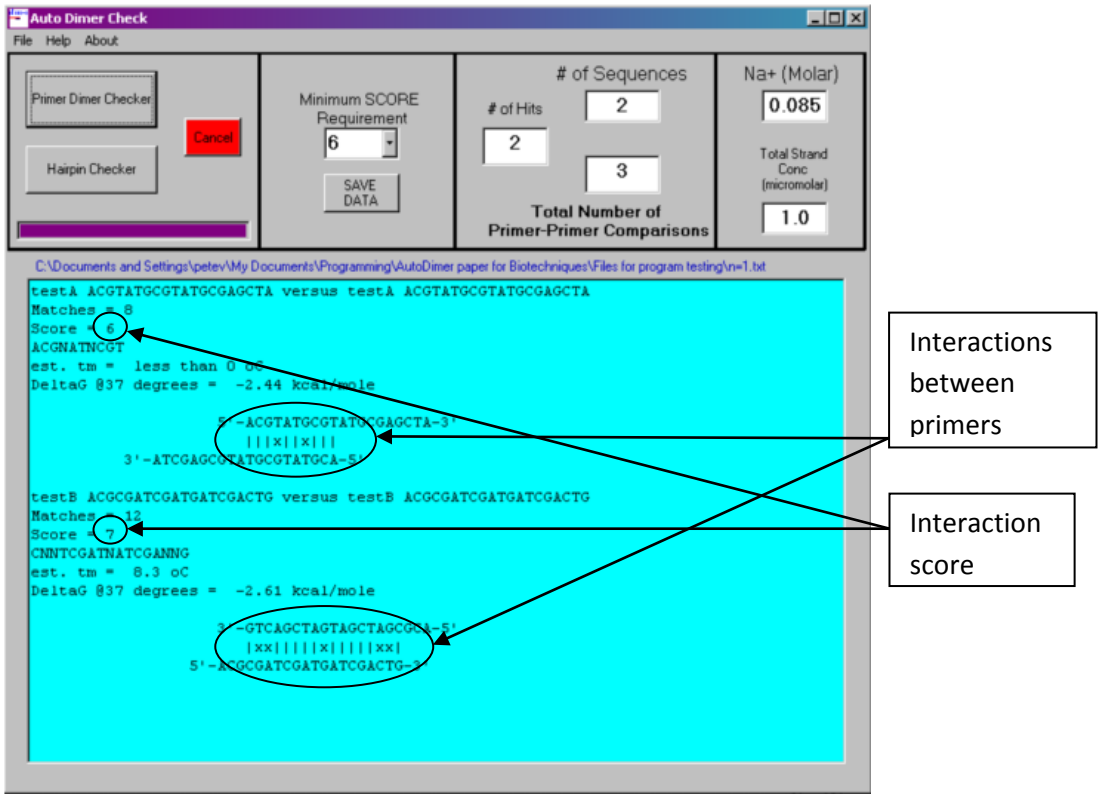


Figure 4.6 AutoDimer v.1 Software.

Input primer sequences are analysed to find potential interactions, shown by aligning the two sequences and joining complementary regions. Lines connecting bases indicate a complementary match, crosses a mismatch. The number of matches minus the number of mismatches gives the score of the interaction, indicated.

Initially, multiplex PCR was carried out with unlabelled primers and products were separated via gel electrophoresis. Figure 4.7 shows the gel electrophoresis results for the multiplex reaction. Three product bands are visible, along with a smeared area of unincorporated primers. It is proposed that only three product bands are visible because the products of HperF.4.1 + HperR.4.1 and HascF.1.4 + Hasc.R.1.2 differ in length by just nine bp. These amplicon sizes are probably too similar to be separated by gel electrophoresis, emphasizing the need for capillary electrophoresis which can separate sequences of just one bp difference in length.

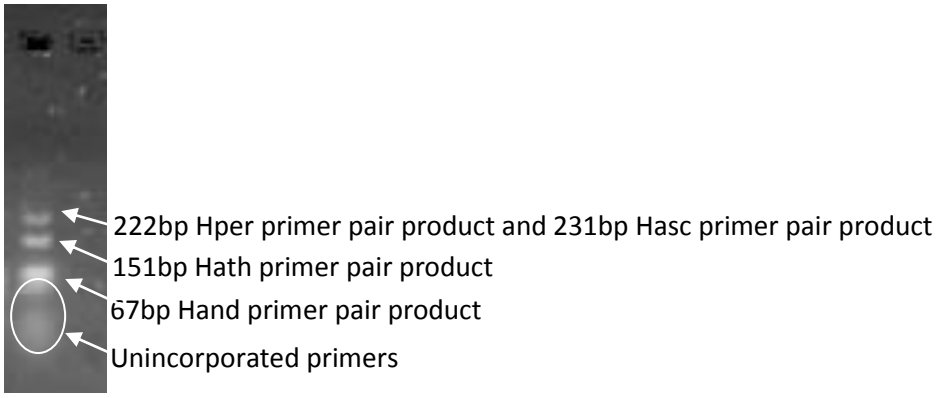


Figure 4.7 Image of gel with products from multiplex reaction.
All selected multiplex primers (Table 4.7) were included, not fluorescently labelled, with the non-kou panel (Table 4.3). This panel contained the target DNA for all four primer pairs included, though three products are visible. The products from HperF.4.1 + HperR.4.1 and HacsF.1.4 + HascR.1.2 differ by only 9bp, this is not separable via conventional electrophoresis so just one band is seen.

The four forward primers were then fluorescently labelled to allow detection of the amplicons via capillary electrophoresis, and again tested against their target DNA and the non-target panels to ensure no reaction differences had occurred due to the fluorescent labelling.

4.3.5 Capillary Electrophoresis

Capillary electrophoresis (CE) is capable of resolving amplicons which differ in length by just one bp, hence its use in DNA sequencing. This enables much higher resolution of the resultant amplicons from this technique, but also confers higher detection sensitivity due to the incorporation of fluorescent labels into the forward primers. It is possible that PCR products which were not visible by conventional gel electrophoresis may have still produced additional unwanted peaks when analysed via CE. To ensure this did not occur, the conventional, or singleplex, PCR reactions were repeated using a temperature gradient run and the products analysed via CE. The results for the *H. athoum* primer pair are shown in Figure 4.8, with both the target and non-ath panel DNA samples. The non-ath panel DNA produced two cross

amplification peaks until the annealing temperature of 64^{o}C, the same temperature required when analysed by conventional gel electrophoresis. This optimisation was conducted for the four primer pairs chosen.

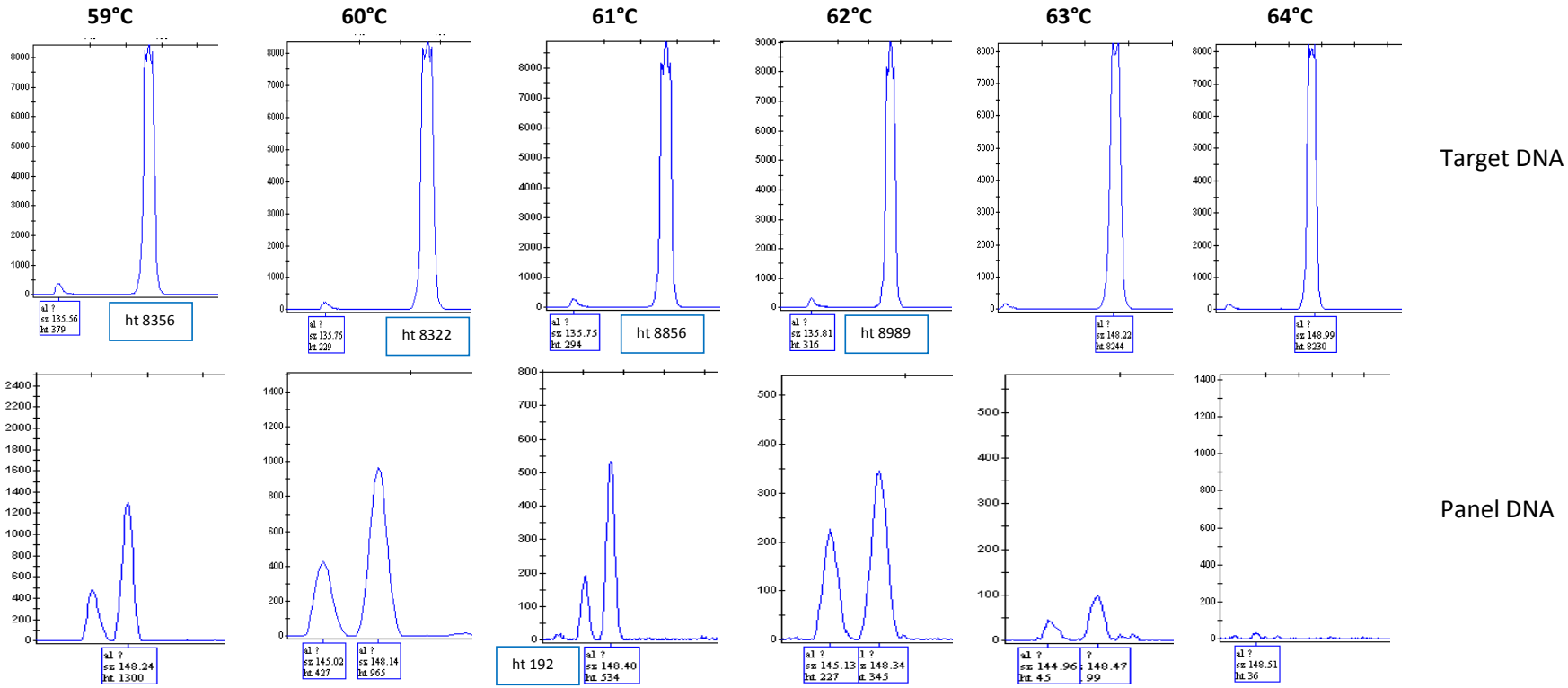


Figure 4.8 Capillary electrophoresis results from singleplex PCR reactions.
The annealing temperature is shown above each column, the top row template DNA was the target *H. athoum* dilution, and the bottom row the non-athoum panel of non-target DNAs. At the lower annealing temperatures cross-amplification occurs creating peaks at 145 and 148 bp amplicon length. The target amplicon is produced at all temperatures, and at 64°C annealing temperature is the only peak in any condition, providing the requirements for the technique.

Multiplex PCR reactions must be optimised separately from the individual PCR reactions. As several reactions occur in one tube, competition for reagents becomes limiting. A highly efficient amplification may out-compete all others, and consume all available dNTPs for instance, preventing other reactions. This can be resolved by altering the concentrations of different primer pairs, to compensate for any lower efficiency pairings. All four selected primer pairings were checked for this and the pair most affected by introduction into the multiplex system was HperfF.4.1 and HperfR.4.1, as shown in Figure 4.9.

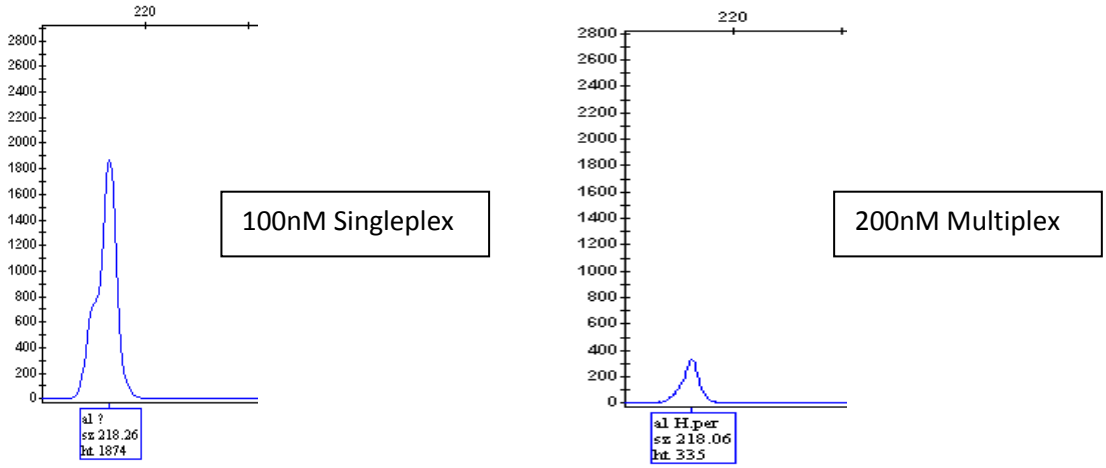


Figure 4.9 Product peaks detected for primers HperfF.4.1 and HperfR.4.1 in singleplex and multiplex reactions.

Although the primer concentration was increased in the multiplex reaction, the resultant peak is still too small to be reliably detected. Based on these results, the concentration of the Hperf primers were further increased in the multiplex reactions to 300nM, and other primer pairs were reduced in concentration, see section 4.2.4 for full details.

The multiplex reaction was conducted at several different primer concentrations to account for the different efficiencies of the primer pairs, the results of which are shown in Figure 4.10. Four grey areas are shown on the electropherogram, these are the 'bins' input by the user. Each bin is 5bp in size and correlates to the size of one of the species specific amplicons. Peaks falling within these boundaries therefore confirm the presence of the target DNA in the sample tested. The example shown in Figure 4.10 is the product of the working multiplex reaction with the target DNA for all four species. A peak is present in each of the four bins and has been 'called' by the software. This is the final working assay, capable of individually identifying the DNA of four different *Hypericum* species within one reaction and one analysis.

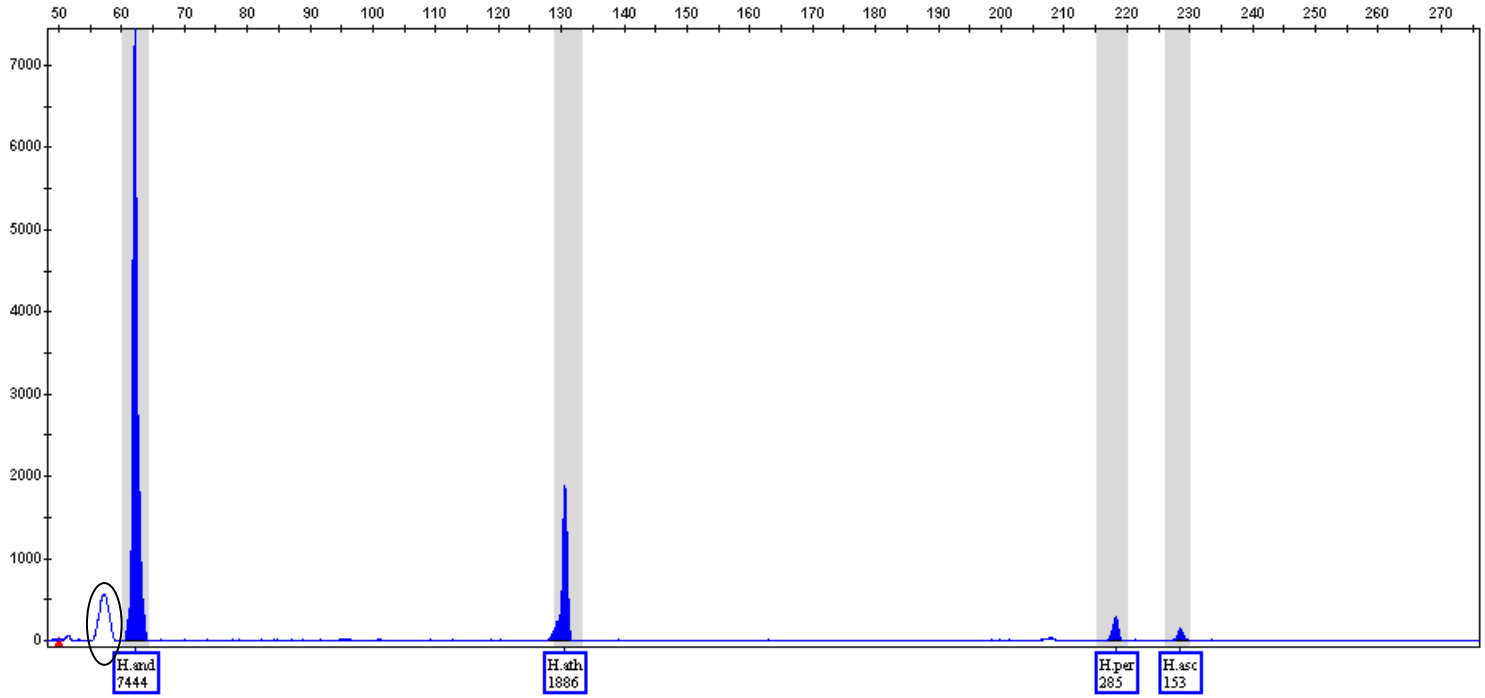


Figure 4.10 PlantID working assay.

Peaks are present for each of the four target species, as indicated by labels. The shaded area the peaks are surrounded by is the 'bin', the size range in which each species specific peak is identified. The multiplex reaction was conducted with all four target DNAs at working concentration. The size differences in the peaks are due to the efficiency of the amplification. Further optimisation could be conducted to ensure equal heights of peaks for all four species. Indicated in a circle is an artefact peak. This peak is present in many samples and in the trace results for four different fluorophores, despite the use of only one in the assay. This peak is likely to have been caused by the size of the Hand peak, fluorescent levels this high can cause 'pull-up' of the baseline due to cross over of spectral readings for different fluorophores within the software.

4.3.6 Conclusion

The design of this assay aimed to identify seven very closely related plant species with the use of one region of nuclear DNA, the nrITS, and this was achieved for four of the target species.

This success was dependent on many factors, one of the most fundamental being the use of AlleleID software. The primer pairs recommended and designed by this software were of extremely high quality, all produced amplicons of the designated size in practice and a large number of candidate primers were found to be species specific within the samples tested. The efficiency of primer design in this example is also due to the quality of the published DNA sequence information, which is a pre-requisite when considering this type of investigation. In future, high quality sequence data should be available for all economically and medicinally important plant species due to the DNA barcoding initiative, enabling the design of this technique to any of these species. The caveat to this is that the identity of the species to be tested for, whether an adulterant or an essential plant for the polyherbal mixture, must be known prior to design of the assay, which in some cases may not be possible

The process of primer design is inevitably followed by empirical testing. This requires template DNA to be available for all species within the design. In this example, the acquisition of sufficient quantities of template DNA proved very difficult, but was circumvented by the use of high fidelity PCR amplicons as the template DNA. This proved an extremely useful tool in assay design, but now necessitates another experimental phase of testing against genomic DNA samples to ensure the same results are obtained.

The empirical testing of the potential primer combinations was labour intensive and time consuming. Despite the efficiency of the AlleleID software to design effective primers, and the use of further software such as AutoDimer to assess multiplex compatibility, each reaction can only be reliably tested in the laboratory and may require optimisation. Optimisation of the multiplex assay is also time consuming, and must be approached with a trial and error based methodology adding one primer pair at a time and correcting any unwanted effects as they occur. Further work on the optimisation of this assay would aim to optimise the multiplex assay to the extent where each peak produced would be of the same intensity if input DNA templates were at an equal concentration. This could then produce a semi-quantitative assay, relative peak heights indicating which DNA is present at the highest and lowest concentrations.

Variation in the sequences used for design confers the possibility for successful species-specific primer design, essential for this assay. An important question presented by this work is whether the nrITS region is variable enough in the genus *Hypericum* to support multiple species identification. More investigation is required to assess whether this or another region, potentially one of the now named barcode regions, is the best platform for assay design (See Section 5.3.5.4).

It is possible that one region of DNA alone will never be sufficiently variable to design species specific primers for any selection of species, and that several regions should always be used in this type of assay design. It follows that the use of more DNA regions will allow more possibilities of unique annealing positions for primers. The use of multiple DNA sequence regions for assay design could also be used to confer greater reliability. If species were identified by the presence of several peaks, the possibility of a false positive result would be greatly reduced. These regions could also be strategically chosen to identify plant species by different genomes, one marker in the nuclear genome, one in the plastid and another in the mitochondrial for instance. This could be achieved using the barcode regions *rbcL* and *matK* in conjunction with the nrITS.

The research in this chapter has provided a proof of concept for this type of assay within medicinal plants, and as such was designed using very closely related plant species as a worst case scenario. The plant species to be identified in genuine cases of misidentification or adulteration are unlikely to be this closely related, and so are likely to contain more DNA sequence variation between them. This variation would enable identification of a much greater number of species per assay, as is the case in the animal identification assay which can identify 18 species (Tobe and Linacre, 2008).

The technique described in this chapter provides a unique and powerful tool in medicinal plant identification in that it fulfils two requirements within one protocol:

- Verification of intended plant species, one or many
- Identification of adulterant plant species

Particular situations which would benefit from the directed development of this type of assay include polyherbal preparations, for which no other technique can confirm the presence of each individual species.

For example, Ayurvedic preparations such as Dashmoola which contain many different plant species, but each will be highly processed making them impossible to identify morphologically. The chemical analysis of such a preparation would contain many compounds from all of the different species, producing a highly complicated data set with no way of isolating which plant any compound had come from.

As discussed in Section 1.3.1, substitution of the raw materials of this preparation is common, and the problems described above led to the development of a DNA based assay to identify one species which should be in the preparation, and two which are often found as adulterants, *Desmodium gangeticum, D. velutinum* and *D. triflorum* respectively. Each of these is identified by an individual PCR reaction, which is then analysed by gel electrophoresis. However, the multiplex PlantID system could potentially identify all ten different species which should be present in the preparation, and test for species which are known to be used as adulterants in one reaction analysed in one procedure.

4.3.6.1 Acknowledgements

The work described in this chapter was funded by a Higher Education Collaboration Grant from the Healthcare and Bioscience iNet, and performed in collaboration with the East Midlands Forensic Pathology Unit, Leicester University.

5 DNA Sequence Analysis

5.1 Introduction

5.1.1 The genus *Hypericum*

The *Hypericum* genus has been intensively studied for over thirty years by Dr Norman K. Robson of the Natural History Museum London, separating and cataloguing the species based on floral and vegetative morphology. The *Hypericum* genus is extremely large, consisting of 469 species separated into 36 taxonomic sections (Table 5.1), making this is one of the largest and most complex genera to have been fully systematically studied (Robson, 2006). The final relationship between the 36 sections of the *Hypericum* genus has recently been published, and is shown in Figure 5.1.

The nrITS region of a selection of *Hypericum* species has been sequenced and studied on two occasions. Park and Kim (2004) sequenced thirty-six *Hypericum* species from eight sections, with exclusive emphasis on the Korean and Japanese evolution of the genus. This study was the first to publish an nrITS based phylogeny, and found the section *Hypericum* separated into four lineages within Korean and Japanese samples (Park and Kim, 2004).

Crockett et al. (2004) investigated fourty-nine species across eleven sections of the genus, and found that the resultant phylogenetic tree separated the species in broad agreement with the section categorisation of Dr NK. Robson, with the advancement of the identification of three Clades; A, B and C (Figure 5.2). Clades A and B cover several sections of the genus, six and four respectively, and contain mainly Old World species. Clade C is made up entirely of species from the section *Myriandra*, no. 20 in Figure 5.1.

The most economically important and well known species within this genus, *H. perforatum*, falls within the taxonomic section *Hypericum*, no. 9 on Figure 5.1. This section is one of six, 9 to 9e, which radiate from section 7 *Roscyna*, making species within these sections the closest relatives of *H. perforatum*. The assays designed in sections 2, 3 and 4 of this thesis have focused on the nine *Hypericum* species for which vouchered DNA samples had been obtained from the Royal Botanic Gardens Kew. These samples are closely related to *H. perforatum*, and spread throughout Clades A and B (Figure 5.2).

Table 5.1 The sectional classification of *Hypericum* proposed by Dr N.K. Robson.

Taken from (Carine, 2010).

Section number	Name	Distribution
1.	*Campylosporus*	Africa, Madagascar, SW Asia
2.	*Psorophytum*	Balearic Islands
3.	*Ascyreia*	S & E Asia, N. Turkey
4.	*Takasagoya*	Taiwan, Phillipines
5.	*Androsaemum*	Mediterranean, W. Europe, Atlantic Is.
6.	*Inodora*	NE Turkey, SW Georgia
6a	*Umbraculoides*	Mexico (Oaxaca)
7	*Roscyna*	NE Asia, E. North America
8	*Bupleuroides*	NE Turkey, SW Georgia
9	*Hypericum*	Northern temperate regions
9a	*Concinna*	N. California
9b	*Graveolentia*	North & Central America
9c	*Sampsonia*	S. Japan, Taiwan, C & S China, Vietnam, Myanmar, Assam
9d	*Elodeoida*	E & S China, Vietnam, Myanmar to Kashmir
9e	*Monanthema*	SW China, Vietnam, Laos, Thailand, Myanmar to Pakistan, S. India, Sri Lanka
10	*Olympia*	S. Balkans, W. Turkey
11	*Campylopus*	NE Aegean
12	*Origanifolia*	Cyprus, Turkey, Georgia
13	*Drosocarpium*	Mediterranean, Balkans, SW Asia
14	*Oligostema*	Europe, NW Africa, Atlantic Is.
15	*Thasia*	NE Aegean
16	*Crossophyllum*	N & W Turkey, Caucasus
17	*Hirtella*	W. Mediterranean to Altai
18	*Taeniocarpium*	Europe to Altai and Iran
19	*Coridium*	W. Mediterranean to Caucasus
20	*Myriandra*	E North America, Caribbean, Bermuda
21	*Webbia*	Atlantic Is.
22	*Arthrophyllum*	S. Turkey, Levant
23	*Triadenioides*	Socotra, Levant, S. Turkey
24	*Heterophylla*	NW Turkey
25	*Adenotrias*	S. Morocco to Levant
26	*Humifusoideum*	New Guinea, SE Asian Is., trop and S. Africa, Madagascar
27	*Adenosepalum*	Atlantic Is., Africa, Mediterranean Europe
28	*Elodes*	W. Europe, Azores
29	*Brathys*	North & South America
30	*Trigynobrathys*	North & South America, Africa, E. Asia, Australasia

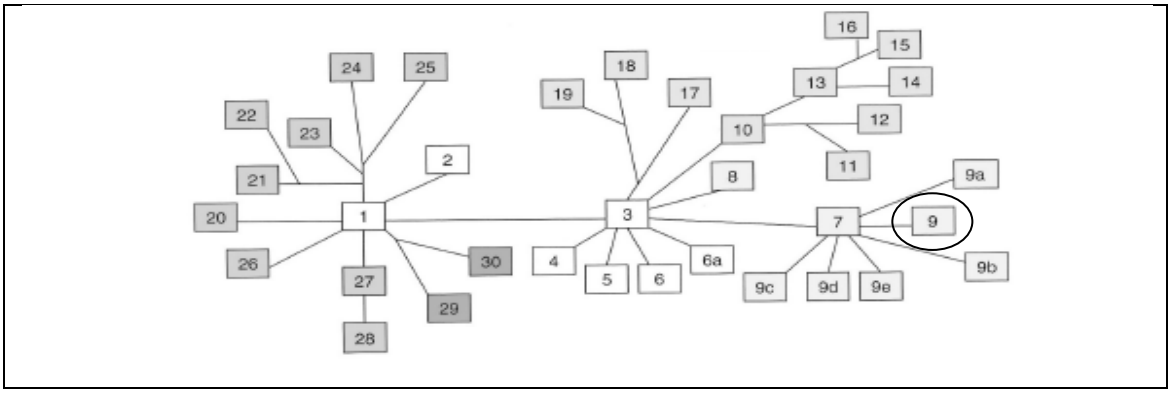


Figure 5.1 The relationship between the different sections of the *Hypericum* genus as defined by Dr N. Robson.

Section *Hypericum*, no. 9, is highlighted; this is the section in which *H. perforatum* is classified. Image taken from (Carine, 2010).

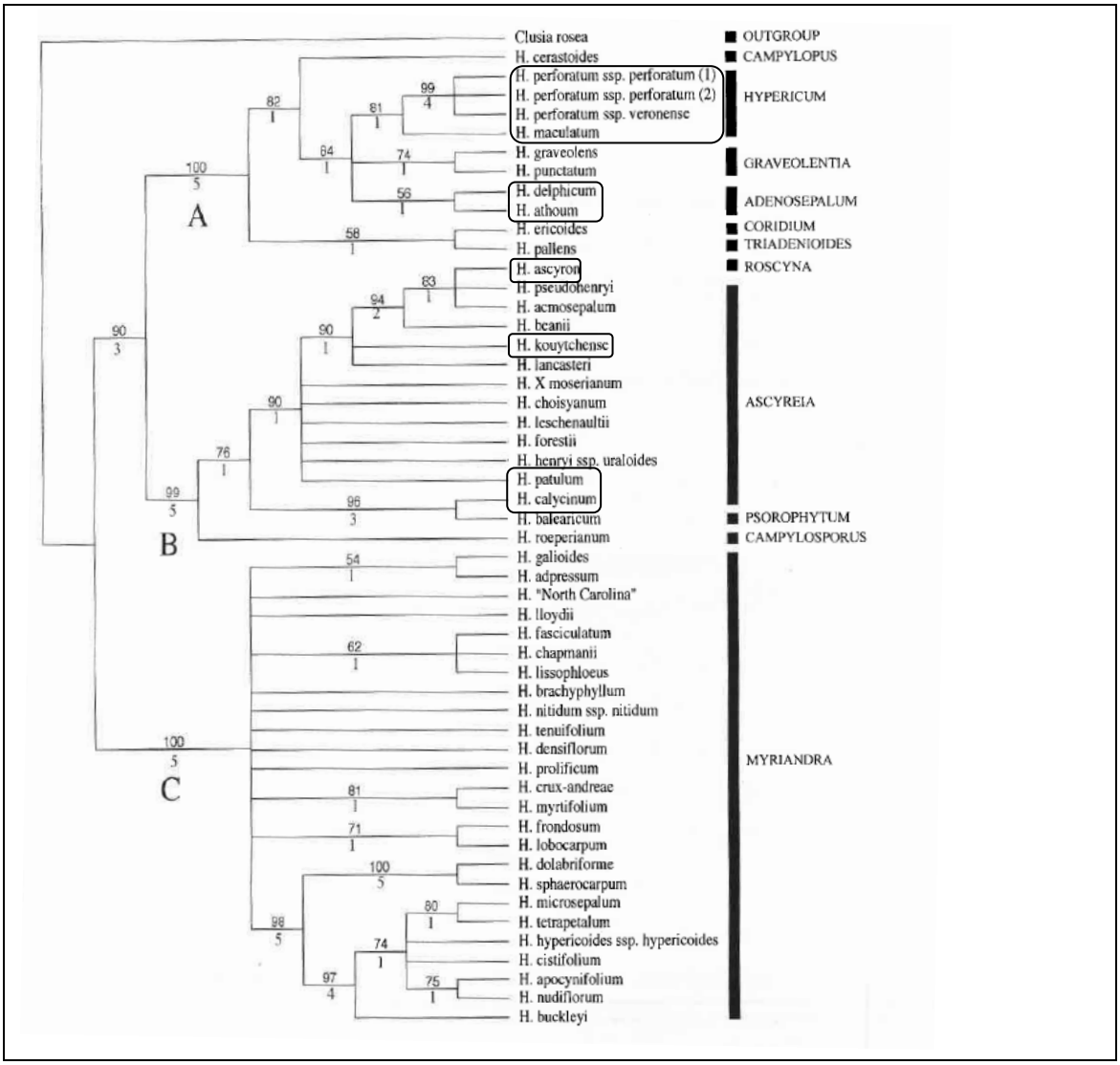


Figure 5.2 The phylogenetic tree produced from the nrITS sequences of 49 *Hypericum* species.

The three clades identified are shown as A, B and C, sectional classifications are indicated by the black bars on the right. The species studied in sections 2, 3 and 4 of this thesis are highlighted, with the exception of *H. androsaemum* which is not included. Image taken from (Crockett et al., 2004) and adapted.

5.1.2 *H. perforatum* and *H. maculatum*

Following publication of the microcode PCR assay to detect *H. perforatum*, a selection of *H. perforatum* and *H. maculatum* samples from different areas of Lithuania was made available for analysis. This presented an opportunity to validate the assay by comparing the response to multiple samples of the target species and one of its nearest relatives (twenty-two and sixteen of each respectively).

The sample set was to be investigated for differences in the composition of essential oils, and collaboration with the De Montfort University Group was suggested in order to explore the possibility of genetic variation correlating with essential oil variability. Previous studies had indicated differences in *H. perforatum* essential oil content which was found not to be due to the environment, and was therefore considered to have genetic causes (Radusiene et al., 2005).

H. maculatum is a species of particular interest when investigating *H. perforatum* as *H. perforatum* is thought to be an allotetraploid which was the result of a hybridisation event between *H. maculatum* and *H. attenuatum* (Robson, 2002). It has also been suggested that *H. perforatum* is in fact an autotetraploid, due to differences in the positions of different genes as detected by fluorescent in situ hybridisation (FISH) (Brutovská et al., 2000). As *H. maculatum* and *H. perforatum* are so closely related, the gold standard for an *H. perforatum* DNA-based identification technique is the ability to reliably and consistently differentiate between the two species.

In the initial testing of the microcode PCR assay, the vouchered *H. maculatum* sample from Kew did not amplify with the primers FO2 and HRI-S (Figure 2.5). This was noted as unusual at the time, as the similarity of the published *H. perforatum* and *H. maculatum* nrITS sequences was the same as *H. delphicum* which did amplify. However, as the two published *H. maculatum* sequences disagreed by one base within the annealing position of FO2 (Figure 2.6) this was thought be due to inconsistencies in the published data.

In order to verify the microcode PCR technique, and begin further investigations, DNA extraction was carried out for all samples, and each was tested using the primers FO2 and HRI-S. Unexpectedly, all of the samples gave a positive result, Figure 5.3.

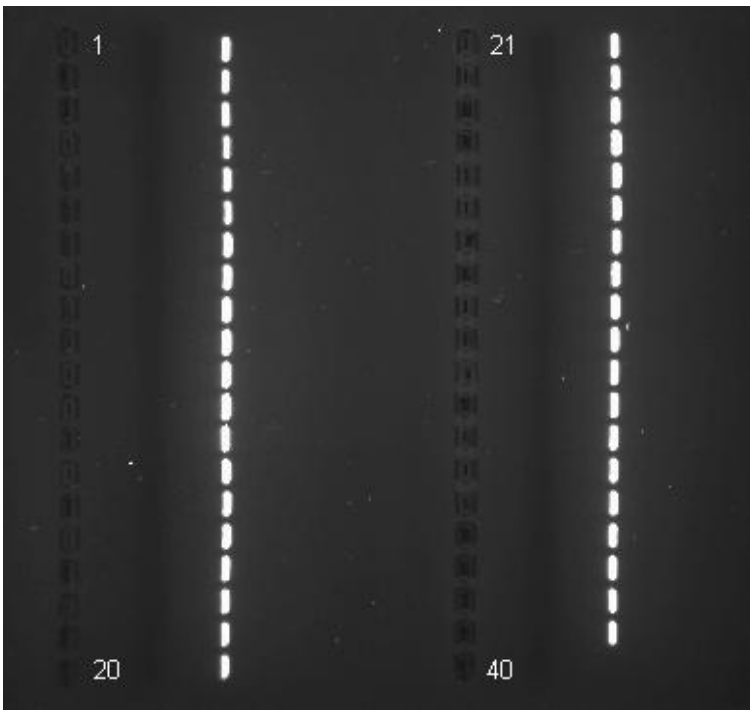


Figure 5.3 Image of gel showing the 'species-specific' product of FO2 and HRI-S with all *H. perforatum* and *H. maculatum* samples from Lithuania.

All samples produced the 85bp product which had previously been shown only to be specific to *H. perforatum*, and did was not produced by the vouchered *H. maculatum* sample from the Royal Botanic Gardens, Kew. Lane 39 was an *H. perforatum* positive control and lane 40 the negative control.

The positive result from all of the Lithuanian samples prompted many questions. There is clearly a DNA sequence difference between the Lithuanian *H. maculatum* samples and the Kew *H. maculatum* sample, so is this the single base difference seen in the published *H. maculatum* nrITS sequences (Figure 2.6)? Is it possible that the *H. perforatum* samples have different nrITS sequences as well as the *H. maculatum* samples? In order to answer these questions the nrITS region of all samples must be sequenced and the results analysed.

5.1.3 Barcode Analysis in *Hypericum*

The results of the multiplex PlantID system indicated that the nrITS was suitably variable to enable the design of species specific primers to four of seven closely related *Hypericum* species (Section 4.3.6). This implies that in order to identify more species using this technique, the nrITS alone may not be sufficient, and another DNA region would be required.

The design of all of the DNA-based identification assays within this work was based on the concept of using the nrITS sequence data as a model for the barcode data which would eventually be available for most, if not all, plant species. During the design of these assays, the *matK* and *rbcL* plastid regions were selected as the plant barcodes. These regions could

reasonably be expected to produce the additional platform required for further DNA-based identification techniques, and aid the development of the current ones.

The chosen barcode regions have been tested in many different species across different genera (1.3.9). However, the literature does not contain a study of these regions within the *Hypericum* genus. Important information concerning the ease of amplification and protocols required to amplify and sequence these regions within *Hypericum* species is therefore unknown.

In addition to this, twenty-three new vouchered specimens were supplied by the Natural History Museum London. The twenty-three species, listed in Table 5.2, include representatives from twelve sections of the genus. As of August 2009, the GenBank sequence database contained no records for the nrITS or any of the barcode regions for these species.

The sample sets described can be divided into two main categories;

- Vouchered samples from thirty-two different *Hypericum* species, one sample per species, with two further samples for the most economically important species *H. perforatum*. Eleven of these were supplied by the Royal Botanic Gardens Kew DNA databank, and twenty-three by the Natural History Museum London.

- A selection of *H. perforatum* and *H. maculatum* samples from different areas of Lithuania, twenty-two and sixteen of each respectively.

These two categories present an opportunity to assess the utility of the barcode regions within *Hypericum*, as they enable both *inter* and *intra* species variation to be measured. In accordance with the recommendations of CBOL, the *trnH-psbA* region should also be sequenced as a 'back-up' barcode (Executive Committee, 2009).

In addition, the nrITS region of the majority of the twenty-three samples from the Natural History Museum had not been sequenced, nor included in the previous phylogenetic studies mentioned (section 5.1.1). The sequencing of this region would therefore uncover the previously unknown phylogenetic relationship of these species which could then be incorporated with the work of Crockett et al. 2004 and compared to the morphology based proposal of Robson (2006).

5.1.4 Aims

This research aims to;

- Amplify and sequence the nrITS region of the NHM samples to assess variation and infer phylogenetic relationships.
- Amplify and sequence the nrITS region of the Lithuanian samples to discover the cause of a false positive result from the microcode PCR test and assess variation.
- Amplify and sequence the barcode regions of *rbcL* and *matK*, and the 'back-up' barcode *trnH-psbA*, in all samples to determine the most informative barcode for use in *Hypericum* species for species identification.
- Analyse the possibility of making phylogenetic inferences from the data of the barcode regions.
- Establish the region most suitable for the design of DNA-based identification assays.

5.2 Materials and Methods

5.2.1 DNA Sample Materials

Vouchered *Hypericum* DNA samples were obtained from The Royal Botanic Gardens Kew, with details as described previously (Section 2.2.2). Twenty-three further vouchered DNA samples were provided by Dr Mark Carine of the Natural History Museum London. The represented species are listed in Table 5.2 - this was termed the NHM set.

Table 5.2 DNA samples provided by the Natural History Museum

Reference Number	*Hypericum* Species	Authority	Section within Genus
A09	*H. bellum*	H.L.Li	*Ascyreia*
A10	*H. klusianum*	Unlisted	*Drosocarpium*
A12	*H. montanum*	L.	*Adenosepalum*
B09	*H. henryi subsp. Henryi*	H.Lév. & Vaniot	*Ascyreia*
D05	*H. epigeum*	Unlisted	*Graveolentia*
D06	*H. marginatum*	Woronow	*Taeniocarpium*
D08	*H. filicaule*	N.Robson	*Monanthema*
D10	*H. coris*	L.	*Coridium*
D11	*H. laxiflorum*	N.Robson	*Origanifolia*
E01	*H. quartianianum*	Unlisted	*Campylosporus*
E05	*H. pseudomaculatum*	Bush	*Graveolentia*
E06	*H. thymifolium*	Sol.	*Taeniocarpium*
E08	*H. maclarenii*	N.Robson	*Campylosporus*
E09	*H. wardianum*	N.Robson	*Campylopus*
F07	*H. elatoides*	Keller	*Ascyreia*
F08	*H. monogynum*	L.	*Ascyreia*
F09	*H. fosteri*	N.Robson	*Origanifolia*
G04	*H. senanensis*	Unlisted	*Hypericum*
G06	*H. confertum*	Choisy	*Taeniocarpium*
G08	*H. wilsonii*	N.Robson	*Takasagoya*
G09	*H. latisepalum*	N.Robson	*Drosocarpium*
H05	*H. elodioides*	Unlisted	*Graveolentia*
H12	*H. rumeliacum*	N.Robson & Strid	*Drosocarpium*

Samples of dried *H. perforatum* L. and *H. maculatum* Crantz plant material were provided by Asta Judzentiene of the Institute of Chemistry, Vilnius, Lithuania. These samples, as listed in Table 5.3, were collected from different parts of Lithuania as shown in Figure 5.4 and Figure 5.5. DNA was extracted from each sample as described previously (Section 2.2.2).

Table 5.3 *Hypericum* samples supplied from Lithuania by Asta Judzentiene.

H. perforatum samples			*H. maculatum* samples		
Sample No.	Habitat	DNA Extraction No.	Sample No.	Habitat	DNA Extraction No.
1	Kalvarija, Jungenai	005	1	Vilnius, Botanical Garden	007
2	Vilkaviskis, Naudsiai	043	2	Vilnius, Botanical Garden	011
3	Skudas, Mosedis	017	3	Vilnius, Botanical Garden	012
4	Anyksciai, Svedasai	026	4	Vilnius, Botanical Garden	018
5	Silute, Silute	016	5	Vilnius, Botanical Garden	030
6	Kaunas, Girionys	038	6	Sirvintos, Ciobiskis	045
7	Svencioniai, Januliskis	044	7	Raseiniai, Lyduvenai	001
8	Trakai, Rykantai	035	8	Raseiniai, Lyduvenai	031
9	Palanga, Sventoji	029	9	Raseiniai, Dubysa	027
10	Klaipeda, Karkle	015	10	Anyksciai, Svedasai	022
11	Kelme, Tytuvenai	032	11	Kretinga, Kretinga	014
12	Svencioniai, Labanoras	041	12	Kaunas, Girionys	034
13	Jurbarkas, Jurbarkas	028	13	Raseiniai, Katauskiai	025
14	Panevezys, Panevezys	024	14	Raseiniai, Gruzdiske	019
15	Kedainiai, Kampai	013	15	Silale, Pajuris	036
16	Raseiniai, Ariogala	037	16	Raseminiai, Girkalnis	023
17	Jurbarkas, Skirsnemune	039			
18	Neringa, Juodkrante	020			
19	Raseiniai, Raseiniai	042			
20	Prienai, Zarstai	040			
21	Panevezys, Ramygala	033			
22	Druskininkai, Latezeris	021			

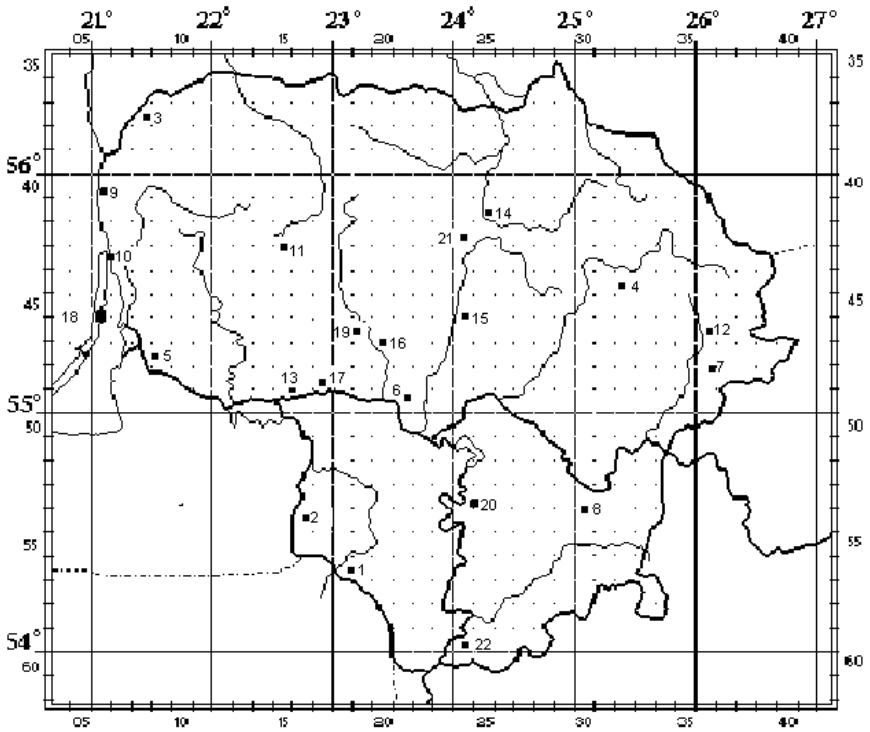


Figure 5.4 Habitats of samples of *H. perforatum* from Lithuania; numbers relate to samples in Table 5.3.

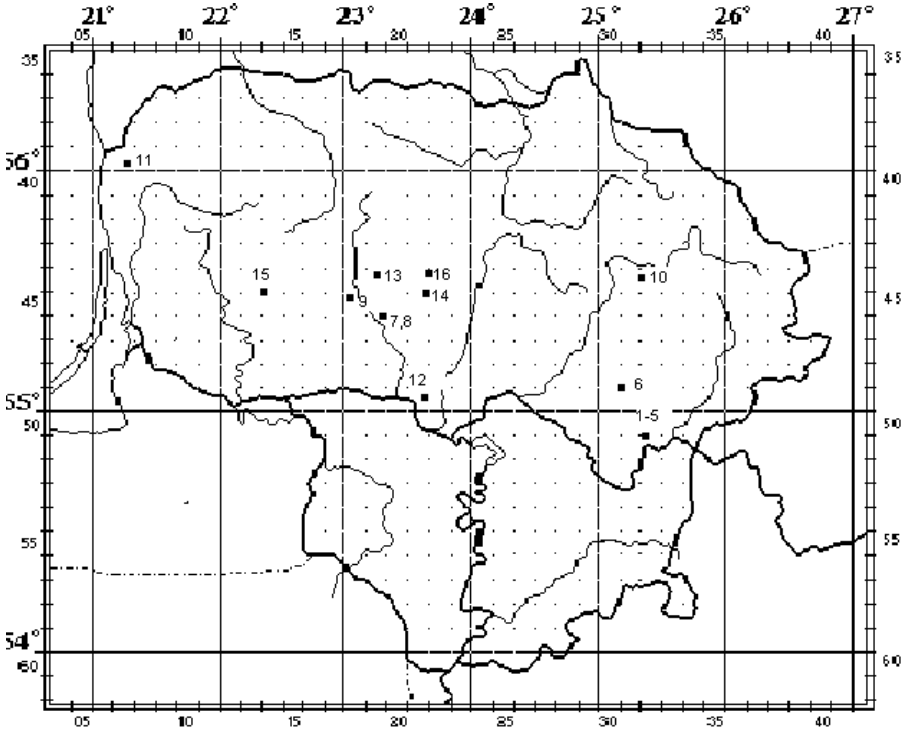


Figure 5.5 Habitats of samples of *H. maculatum* from Lithuania; numbers relate to samples Table 5.3

5.2.2 DNA sequencing

5.2.2.1 Preliminary PCR

Initial PCR amplifications for DNA sequencing were conducted personally and with a Research Assistant, Sarah Smith, under supervision, as described in section 2.2.4, with the following alterations;

PCR primers used for the nrITS region were ITS1 and ITS4 as described in section 2.2.4. The other regions were amplified with the following primers:

trnH-psbA,	trnHf	(5′-CGCGCATGGTGGATTCACAATCC-3′)
	psbA3'f	(5′-GTTATGCATGAACGTAATGCTC-3′) (Kress et al., 2005)
rbcL,	rbcLa_f	(5′-ATGTCACCACAAACAGAAAC-3′)
	rbcLa_rev	(5′-GTAAAATCAAGTCCACCRCG-3′) (Lahaye et al., 2008)
matK,	390F	(5′-CGATCTATTCATTCAATATTTC-3′)
	1326R	(5′-TCTAGCACACGAAAGTCGAAGT-3′) (Cuenoud et al., 2002)
	2.1	(5′-CCTATCCATCTGGAAATCTTAG-3′)
	2.1a	(5′-ATCCATCTGGAAATCTTAGTTC-3′)
	X	(5′-TAATTTACGATCAATTCATTC-3′)
	5	(5′-GTTCTAGCACAAGAAAGTCG-3′)
	3.2	(5′-CTTCCTCTGTAAAGAATTC-3′)
	3F_KIM f	(5′-CGTACAGTACTTTTGTGTTTACGAG-3′)
	1R_KIM r	(5′-ACCCAGTCCATCTGGAAATCTTGGTTC-3′)

PCR cycling parameters for each region were as follows:

nrITS; as described in section 2.2.4.

trnH-psbA; 5min at 95°C initial denaturation step, 35 cycles consisting of 1min at 95°C, 30s at touchdown temperature and 1min at 72°C, final extension period of 7min at 72°C. Touchdown temperature began at 58°C, reduced by 1°C per cycle until 48°C, then continued at 48°C for the remainder of the programme.

rbcL; 5min 95°C initial denaturation step, 35 cycles consisting of 30s at 95°C, 20s at 52°C and 50s at 72°C, with a final extension period of 5min at 72°C.

matK; Initial 'touch-up' programme, 5min 94°C initial denaturation step, 5 cycles consisting of 30s at 94°C, 40s at 44°C and 40s at 72°C, followed by 30 cycles consisting of 30s at 94°C, 40s at 46°C and 40s at 72°C, with a final extension period of 3min at 72°C. The second amplification contained 2µL of the initial PCR product diluted 1:200 as the DNA template.

Second *matK* programme: 5min 94°C initial denaturation step, 35 cycles consisting of 30s at 95°C, 20s at 46°C and 40s at 72°C, with a final extension period of 3min at 72°C.

In all cases, samples without template DNA were used as controls. PCR products were run on 2% (w/v) agarose, 0.5 X TBE gels with 2µL SYBRsafe™ (Invitrogen, Carlsbad, CA, USA) DNA stain at 90V for ~30min and analysed in a Bio-Rad (Bio-Rad, Richmond, CA, USA) Illuminator with ChemiDocXRS Camera and Quantity One software to ensure single banding.

At this point, samples were either sent to Macrogen Europe for sequencing or sequenced in-house.

The sequencing primers used for each region were as follows:

nrITS; forward nITS1 (5′-ACCTGCGGAAGGATCATTGTCGA-3′)

rbcL; forward and reverse as in the preliminary PCR reaction

trnH-psbA; forward HTPSF (5′-AGCTGCTATCCAAGTTCCATC-3′)

second forward HTPSF2 (5′-TCAATCAACACGTCATTGTATCA-3′)

reverse HTPSR (5′-CCAAAAATCTCGGCATGAAT-3′)

matK; HPmK 446F (5′-TTCAAACCCTTCGGTACTGG-3′)

HPmK 446R (5′-CCAGTACCGAAGGGTTTGAA-3′)

HPmK 949R	(5′-TTGGTTGAACCCACACAAAA-3′)
HPmK 949F	(5′-TTTTGTGTGGGTTCAACCAA-3′)
HPmK 1F	(5′-ATGCGGGAAGAGGAATATCA-3′)

The target *matK* amplicons were excised from the electrophoresis gel with Geneclean® II Kit (Qbiogene Inc., Carlsbad USA) following manufacturers' instructions.

5.2.2.2 In-house Sequencing

In-house sequencing was performed personally and by a research assistant, Sarah Smith, under supervision. Preliminary PCR reactions were purified using QuickStep™ 2 PCR Purification Kit (EdgeBio, Maryland, USA) and the DNA quantified using a Qubit® Fluorometer and Quant-iT™ dsDNA BR Assay Kit (Invitrogen, Carlsbad, CA, USA).

Cycle sequencing reactions, final volume 20µL, were conducted in 0.2mL polypropylene tubes using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA). Reactions consisted of Ready Reaction Pre-mix (2.5X) (ABI), BigDye Sequencing Buffer (5X) (ABI), sequencing primer (3.2pM) (VHBio, Gateshead UK or IDT, Iowa USA) template PCR product (5-20ng) and nuclease-free water. Reactions without template PCR product were used as controls.

The Applied Biosystems GeneAmp PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA) was used with programme; 1min at 96^{o}C initial denaturation, 25 cycles consisting of 10s at 96^{o}C, 5s at 50^{o}C, 4min at 60^{o}C. Extension products were purified using Performa® DTR Gel Filtration Cartridges (EdgeBio, Maryland, USA), 10µL Hi Di Formamide was added and the sample thoroughly vortexed.

Products were analysed on the ABI Prism™ 310 Genetic Analyzer (Applied Biosystems, Foster City, CA), using a 47cm capillary and Performance Optimised Polymer 6 (Applied Biosystems, Foster City, CA). The run module used consisted of a 30s injection at 2.0 kV, followed by electrophoresis running at 50^{o}C and 15kV for 36min.

Sequence Analysis 5.2 (Applied Biosystems, Foster City, CA) software was used to collect data, with Basecaller 310POP6, to create the output AB1 file.

5.2.2.3 Sequence Analysis

ITS and *rbcL* sequence data were analysed using both the CLC Main and Genomics Workbenches (CLC bio, Massachusetts USA). Data from the *trnH-psbA* region were analysed using CodonCode Aligner™ (CodonCode, Massachusetts USA) software.

Sequences were trimmed using default settings, and contigs assembled with automated, vote based, conflict resolution. Manual adjustments based on *rbcL* coding sequences were conducted. Sequences were aligned using the ClustalW program (Chenna et al., 2003) and phylogenetic trees were created using the Jalview software (Waterhouse et al., 2009) both available on the European Bioinformatics website, www.ebi.ac.uk. All settings were default. Distance trees were created using Neighbour Joining clustering, and based on average distance as measured by % identity.

Sequences used in multiple alignments were from GenBank, Accession numbers listed in Appendix Section 8.1.1.

5.3 Results and Discussion

5.3.1 nrITS region

The nrITS regions have been widely sequenced within Hypericum species, with 91 sequences available on GenBank as of July 2007 (listed in Appendix Section 8.1.1). The majority of these sequences were contributed by Crockett et al. (Crockett et al., 2004), who found them to be capable of differentiating 50 *Hypericum* taxa. Due to this the published sequence data for the Kew sample set was used, and the NHM set sequenced with analysis quality data produced for 15 samples (65%). For the Lithuanian samples, analysis quality sequence data was obtained for 12 *H. perforatum* (54%) and 12 *H. maculatum* (75%), 11 samples did not produce sufficient quality data for analysis, 7 *H. perforatum* and 4 *H. maculatum*. Amplification of the nrITS region was straight forward with no significant optimisation required. A gel of the PCR products is shown in Figure 5.6.

The sequence data derived from this region was of a slightly lower than optimal quality, with some background peaks present. An example of this data is shown in Figure 5.7. However, the secondary peaks rarely affected base calling and constructing contigs from three or four sequence reads produced a consistent sequence for each sample with very few base call conflicts.

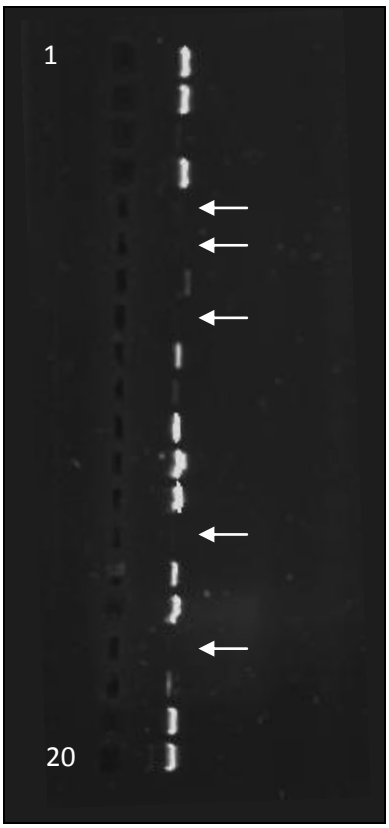


Figure 5.6 Image of a gel with the nrITS amplifications from twenty of the NHM DNA samples.

Five of the reactions failed to produce a sufficiently sized band of product, indicated by arrows.

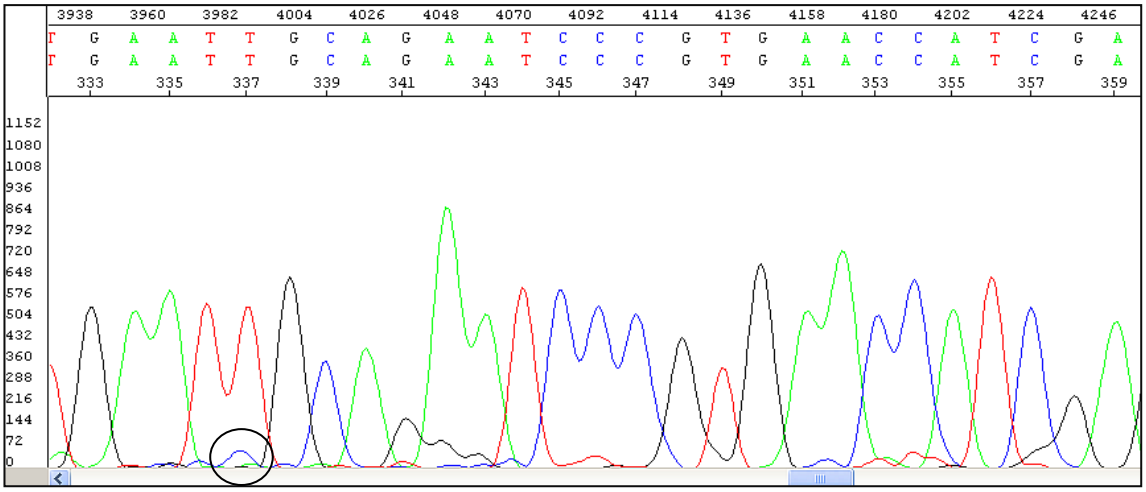


Figure 5.7 Section of the electropherogram from the capillary electrophoresis of an in-house nrITS cycle sequencing reaction with nITS1 primer.

Output direct from ABI Sequence Analysis 5.2 software. A small secondary peak is indicated by a circle, this peak is small enough to be dismissed and has not affected the base calling of the peak as a T.

5.3.1.1 Vouchered samples

The nrITS region has previously been shown to be capable of resolving *Hypericum* species, though this becomes more difficult the more closely related the species are (Crockett et al., 2004). The nrITS phylogenetic analysis of Crockett et al. (2004) grouped the analysed species into three main lineages, Clades A, B and C. Clades A and B mainly consisted of Old World species, whereas Clade C was entirely made up of species from the *Hypericum* genus section *Myriandra*, distributed in East North America, the Caribbean and Bermuda.

The nrITS sequences from the NHM sample set were aligned with the published sequences for the species within the Kew sample set (the full alignment is shown in the Additional Appendix). The distance tree created using this alignment shows agreement with the Clades represented within the sample set (Figure 5.8). Within this sample group there are no accessions from either Clade C, or from the section *Myriandra*. Clade A contains nine species, four of which were included in the original study. Of the five other species present, two are from sections represented by different species in the Crockett study, *H. pseudomaculatum* and *H. montanum* from *Graveolentia* and *Adenosepalum* respectively, and fall within the same clade. The final three species are from sections which were not represented previously, *H. confertum* and *H. thymifolium* from section *Taeniocarpium* and *H. klusianum* from section *Drosocarpium*. These are both Old World sections, supporting and adding to the previous findings.

Clade B contains thirteen species (Figure 5.8) four of which were included in the Crockett study. Six of the species in this Clade are from the section *Ascyreia* (number 3), though these are separated within the Clade. Three sections within this Clade had not previously been studied by nrITS: *Takasagoya* distributed in Taiwan and the Philippines; *Origanifolia* distributed in Cyprus, Turkey and Georgia; and *Drosocarpium* distributed in the Mediterranean, Balkans and S W Asia. A representative from *Drosocarpium* is also present in Clade A, showing that the nrITS data do not correspond exactly to the sectional assignment of each species. Also, *H. wardianum* from section *Campylopus* is in Clade B here, whereas Crockett et al. (2004) placed a sample from this section into Clade A, though this was a different species.

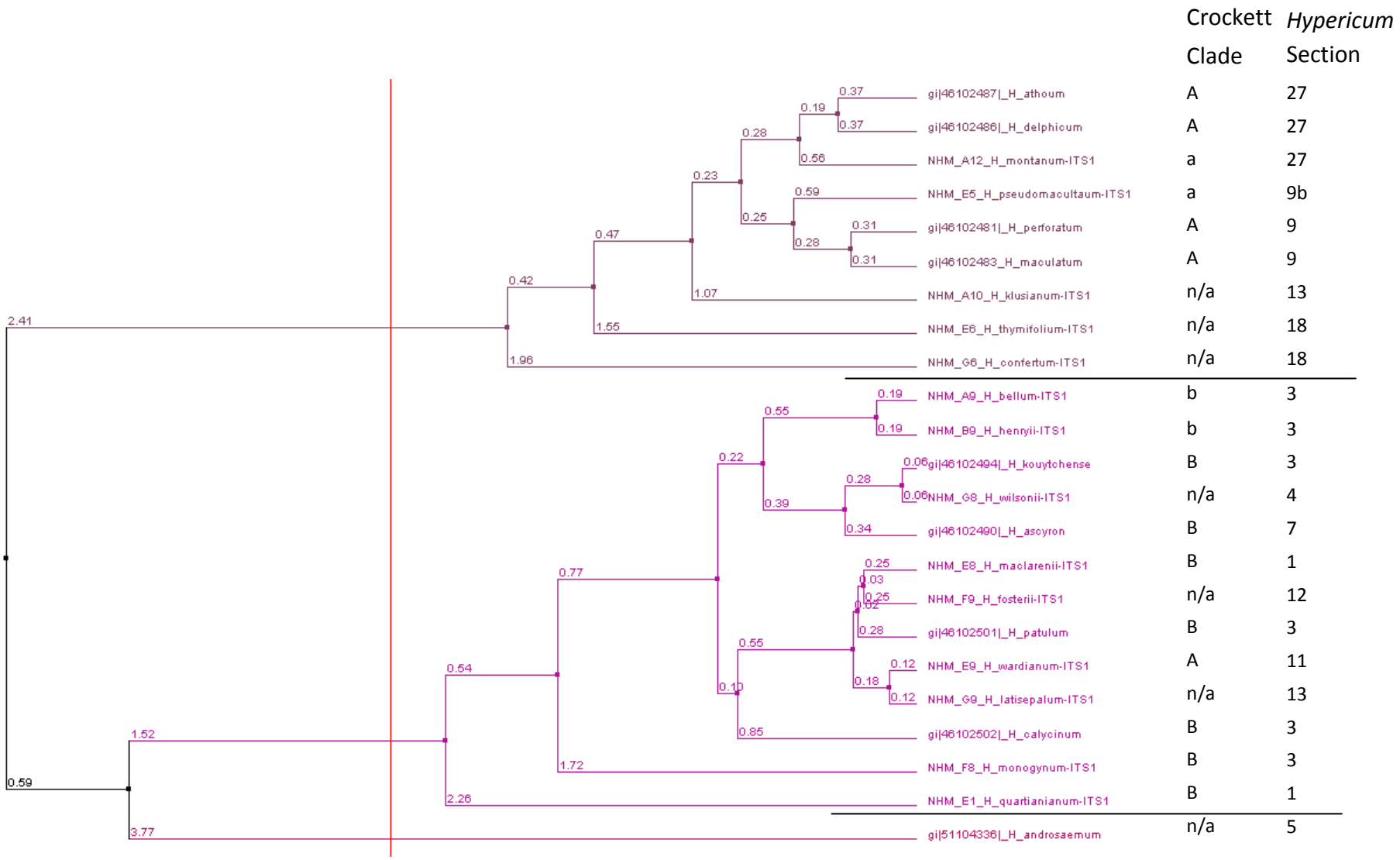


Figure 5.8 Average distance tree created using percent identity for the nrITS region of published *Hypericum* sequences and data produced from the NHM sample set. Shown on the right is the Clade each sample belongs to based on Crockett et al. 2004: Capital letters; this species was in the study, lower case; this section was in the study, and n/a; neither the species nor the section was included. The *Hypericum* genus section number is shown on the far right, relating to Table 5.1.

5.3.1.2 Lithuanian samples

The samples supplied by Asta Judzentiene provided nrITS sequence data for 12 *H. perforatum* and 16 *H. maculatum* samples. In addition to identifying the cause of the unexpected positive results for the *H. maculatum* samples with the *H. perforatum* specific microcode PCR assay, this gave an opportunity to assess the intraspecific variation within this target DNA region, which should ideally be very low in order to meet the requirements of a barcode and to enable species identification.

A multiple alignment of all the *H. perforatum* sequences showed extremely high conservation, with only 22 base differences in total, approximately 0.28% of the total number of base calls. (The full alignment is in the Additional Appendix). Those few base differences that did occur were considered not to be reliable and were attributed to the cycle sequencing chemistry, for instance when they occur close to runs of mononucleotide repeats which can affect base calling and the error rate of *Taq* polymerase (Figure 5.9).

The *H. maculatum* sequences showed a similar degree of conservation, with a total of 27 base differences across all samples, approximately 0.34% of the total number of base calls (the full alignment is in the Additional Appendix). The majority of these base differences are at the very beginning and end of the sequencing read, 12 at the 5′ end (44.4%) and 8 at the 3′ (29.6%), so are likely to result from the unreliability of base calls in these regions rather than representing genuine sequence differences. The other significant proportion of base differences were all from one sample, number 011 (14.8%), (Figure 5.10).

Such high conservation suggests that the nrITS region is not as variable in *Hypericum* as in other genera, and does not present the complications of length variability described in other species (Sass et al., 2007).

Most strikingly, the *H. maculatum* samples show 100% sequence similarity to the annealing position of the microcode PCR assay primer FO2. All but one of the *H. perforatum* samples also show 100% similarity to the primer sequence, the exception being sample 040 which has an additional uncalled base within the region (Figure 5.9). As this base is uncalled it remains questionable, and the electropherogram trace does not resolve this base. For this reason, the base addition in the sequence read is considered to result from the poor quality of the sample.

Equally surprisingly, the annealing position of the microcode assay primer HRI-S is 100% matched by only three of the *H. perforatum* samples. The complement to the beginning of the

primer is the sequence CTCCTT which is perfectly matched by just three *H. perforatum* samples, in the remainder of the *H. perforatum* samples and all of the *H. maculatum* samples the region is CTTCCTT (Figure 5.9 and Figure 5.10). Despite this sequence difference in the penultimate base of the annealing position of HRI-S, all samples gave a positive result producing the target amplicon. Had this sequence difference occurred in the final base (C) it is likely that no product would have been formed.

The sequence polymorphism in the HRI-S sequence was then compared to the available published sequence data. As of 14/05/2009, eight *H. perforatum* nrITS sequences were available on GenBank, though four of these were confined to the coding regions so were not used for analysis. The four remaining sequences, two from Crockett, (Crockett et al., 2004), and two from other unpublished sources with lead authors named as Potter and Kersten respectively, were aligned and had 8 or 11 base differences in total. This is similar to the frequency of base differences in the Lithuanian samples at approximately 0.38% (full alignment in the Additional Appendix). The exact number of base differences depends on a single nucleotide which is called in one sequence as a C, and in the other three as a Y (C or T). The two Crockett sequences were identical and, in most instances of base differences, aligned with the Kersten sequence (Figure 5.11).

A single base difference occurs in the annealing positions of both FO2 and HRI-S primer within the 'Potter' sequence (Figure 5.11). The change in the DNA sequence at the FO2 annealing position is a G to C in the middle of the sequence. Due to the position of the sequence difference, it would be unlikely to prevent amplification, so a sample with the 'Potter' sequence would be likely to still give a positive result with the microcode test. The sequence difference at the HRI-S annealing position is the same as that in the majority of the Lithuanian samples, with an additional T. The published *H. maculatum* sequences also have this extra T. This would suggest that the Kew *H. maculatum* sample used to design the microcode assay contains different sequence variation which is not represented at all in the literature, as it produced a negative result.

It is worth noting that the design of the PCR assay was to microcodes within the nrITS region. These regions are identified based on their variability, enabling distinction between closely related species. It is not unexpected, therefore, that these regions show variability as sample set numbers are increased. It does indicate that multiple markers are likely to be required to guard against further sequence differences causing misleading results.

Figure 5.9 Section of the multiple alignment of the Lithuanian *H. perforatum* nrITS sequences.

The entire alignment contains 22 base differences across all samples. Six base differences are highlighted in blue. The annealing position for the primer FO2 is shown in red, and the beginning of the HRI-S annealing position in green, this covers the only sequence variation seen in the annealing position, three sequence differences within the HRI-S annealing position are shown, these are the only samples which show 100% sequence complementarity to the primer.

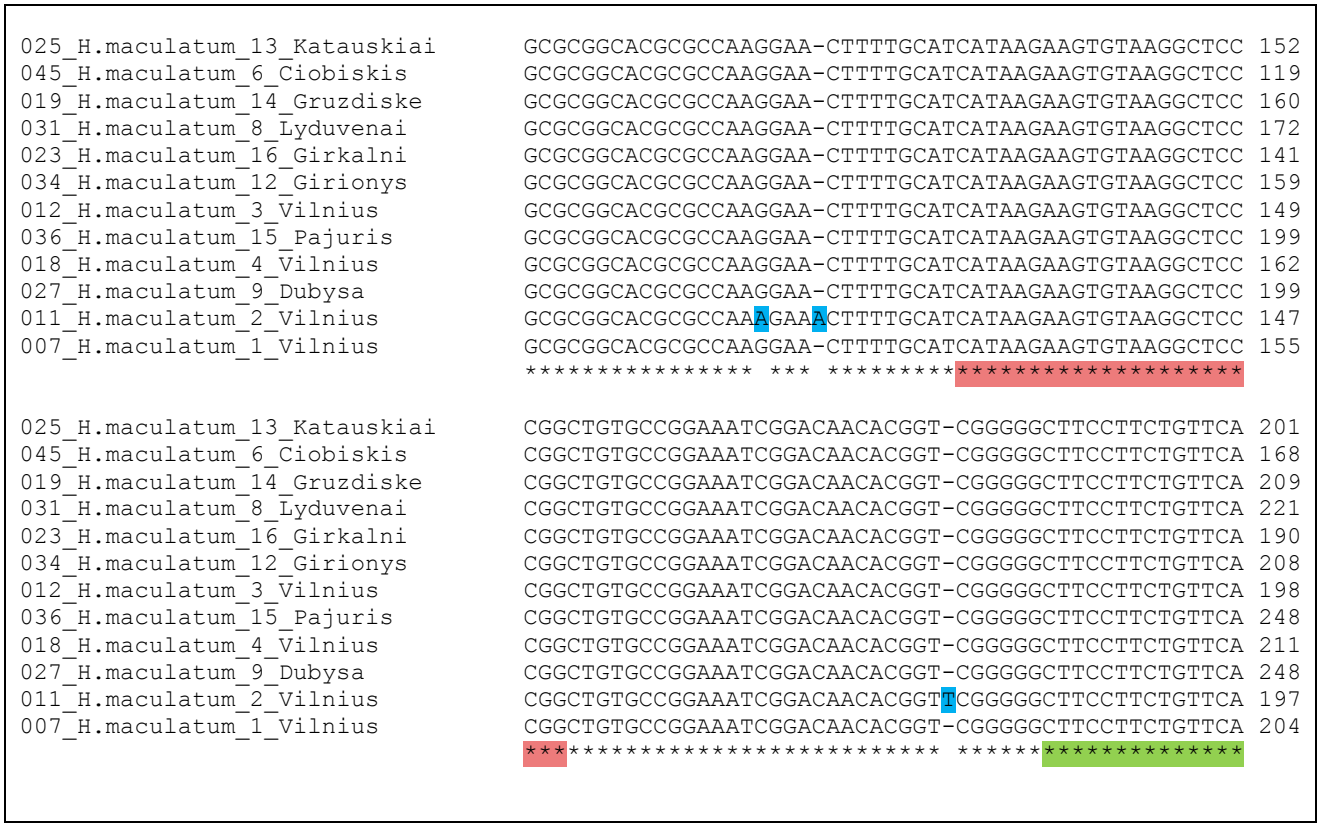


Figure 5.10 Section of the multiple alignment of the Lithuanian *H. maculatum* nrITS sequences.

The entire alignment contains 27 base differences across all samples. Three base differences are highlighted, all occurring in the same sample, 011, which is of slightly lower quality than other samples. The annealing position for the primer FO2 is shown in red, and the beginning of the HRI-S annealing position in green, this covers the only sequence variation seen in the annealing position.

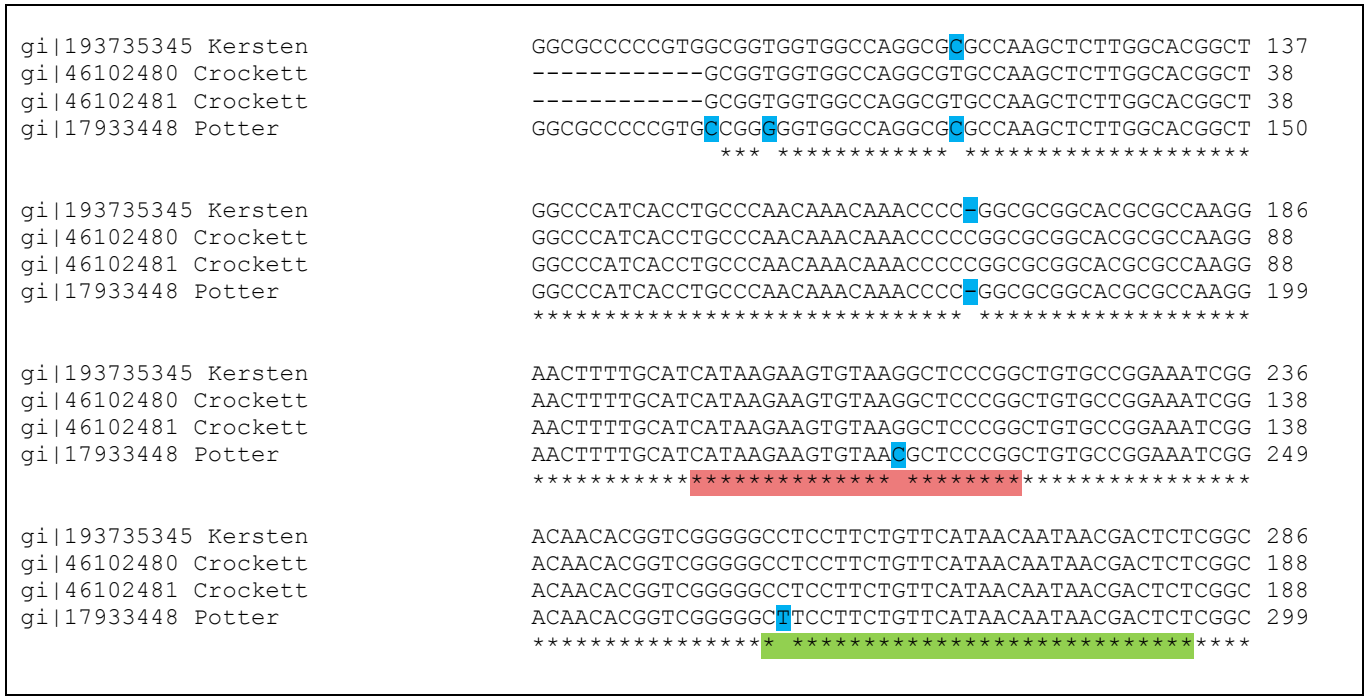


Figure 5.11 Section of the multiple alignment of the four published *H. perforatum* nrITS sequences.

Eight base differences are indicated in blue, four of which occur in the Potter sequence, and four in both the Potter and the Kersten sequence. The Crockett sequences align throughout the sequence and in total there are only 8 to 11 base differences between the four. The annealing position for the primer FO2 is shown in red, and the HRI-S annealing position in green, the sequence difference in the Potter sequence at this point matches the Lithuanian samples.

The Lithuanian samples sequences were aligned with one of the published *H. maculatum* sequences and three of the four published *H. perforatum* sequences (Figure 5.12) (full alignment in the Additional Appendix). Only one *H. maculatum* sequence was used for comparison as the two sequences available differ by just one base. This base is indicated in the alignment. One of the Crockett *H. perforatum* sequences was omitted as the two match each other entirely, so no further information is gained by including both of these sequences.

In total 27 base differences were found across all samples, 0.19% of the total number of base calls. In addition to the base differences discussed in section 5.3.1.2, four further base differences are shown in Figure 5.12. These are an additional base in two samples, 011 and 040. These samples are responsible for the majority of the base difference in the alignment and the sequence data are considered to be of a lower quality. A base addition is shared between the Crockett *H. perforatum* sequence and one *H. perforatum* sample, 024, although the base call is different, C and A respectively. It is possible that this base addition is due to the mononucleotide repeats present on either side, since these bases are each an elongation of one or the other repeat.

Thus, the reason for the positive result of the microcode assay with the Lithuanian *H. maculatum* samples is the high degree of sequence similarity shown between these samples. The sequencing and analysis of the nrITS region shows that the two species represented in the Lithuanian samples set cannot be separated from each other using this region alone.

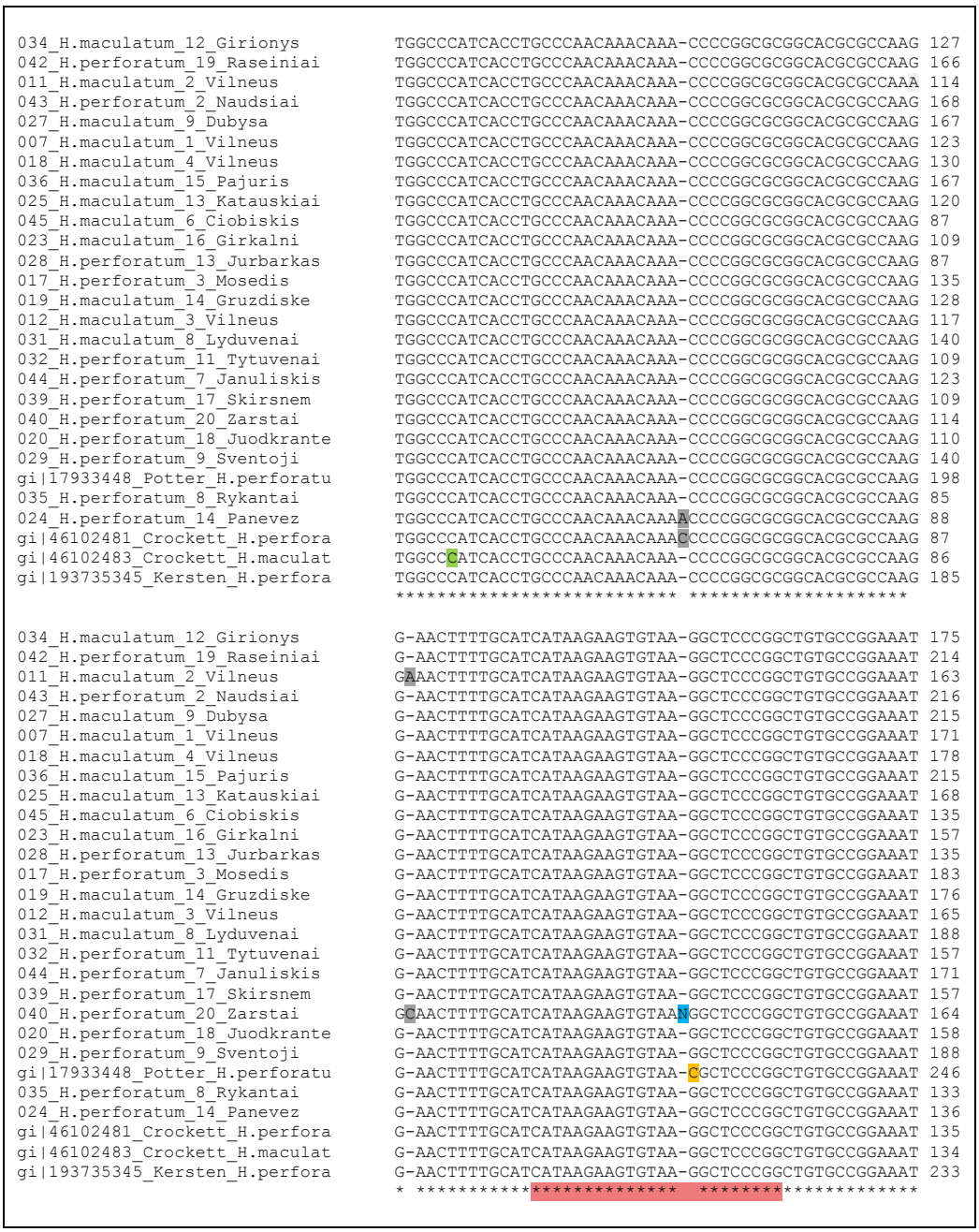


Figure 5.12 Section of the multiple alignment of the Lithuanian samples, plus three published *H. perforatum* and one published *H. maculatum* nrITS region sequences.

The annealing position of microcode assay primer FO2 is shown in red. The only two sequence differences present in this region are the Potter sequence C and the ambiguous N base call, in orange and blue respectively. Indicated in green is the base difference found between the published *H. maculatum* sequences (C to G), which is not shared by any other sample in this set. Four further base differences are shown, an additional base in samples 011 and 040. These samples are responsible for the majority of the base differences in the alignment and are thought to be of lower quality. A base addition is shared between the Crockett *H. perforatum* sequence and one *H. perforatum* sample, 024 although the base call is different. If this base addition is accurate, it is not informative as it is present in so few samples. Overall, the sequences show extremely high conservation.

5.3.2 *rbcL* coding region

Amplification of the *rbcL* region was straight forward for all sample sets and did not require significant optimisation. Analysis quality data was produced for 100% of the Kew samples, 39% of the NHM set, and 100% of the Lithuanian samples. Of the NHM samples which did not produce sequence data, six were in common with the nrITS region, representing 26% of the sample set.

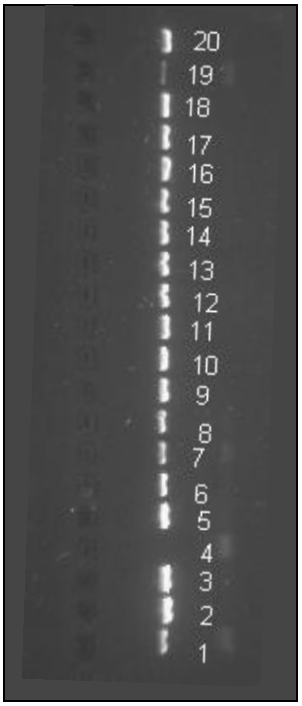


Figure 5.13 Image of gel with *rbcL* products for twenty of the Lithuanian DNA samples.

The *rbcL* region produced very high quality sequence data with few base calling conflicts and secondary peaks; an example is shown in Figure 5.14. Nonetheless two reads, one forward and one reverse, were assembled into contigs to produce a consensus sequence for each sample.

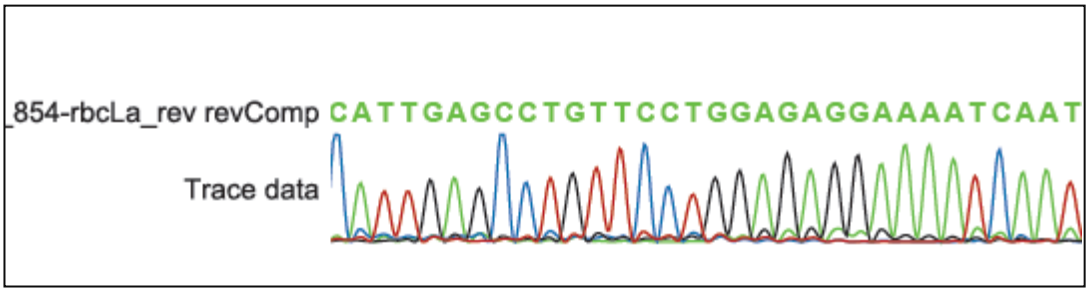


Figure 5.14 Section of the electropherogram from an in-house reverse *rbcL* cycle sequencing reaction as depicted in CLC Main Workbench software.

5.3.2.1 Vouchered Samples

The *rbcL* region codes for ribulose-bisphosphate carboxylase, RuBisCo, which enables an extra step of DNA sequence analysis based on the resultant amino acid (AA) sequence. The consensus sequences produced for each of the vouchered samples were translated into AA sequences which were then aligned with the published *Hypericum* protein sequence (Figure 5.15, full alignment in the Additional Appendix). This enabled the identification of AA sequence differences, which could then be used to work back to the particular nucleotide base differences which had caused them (Figure 5.16, full alignment in the Additional Appendix). These were then assessed as to the probability of the AA change, and the certainty of the base call which had caused it.

Figure 5.15 Section of the amino acid sequence multiple alignment for the vouchered samples and the published *H. perforatum* sequence, Accession no. gi 17135960.

A very unlikely substitution is indicated in orange; a termination is substituted for W - Tryptophan. A possible substitution is indicated in blue and yellow; Y – Tyrosine for F – Phenylalanine.

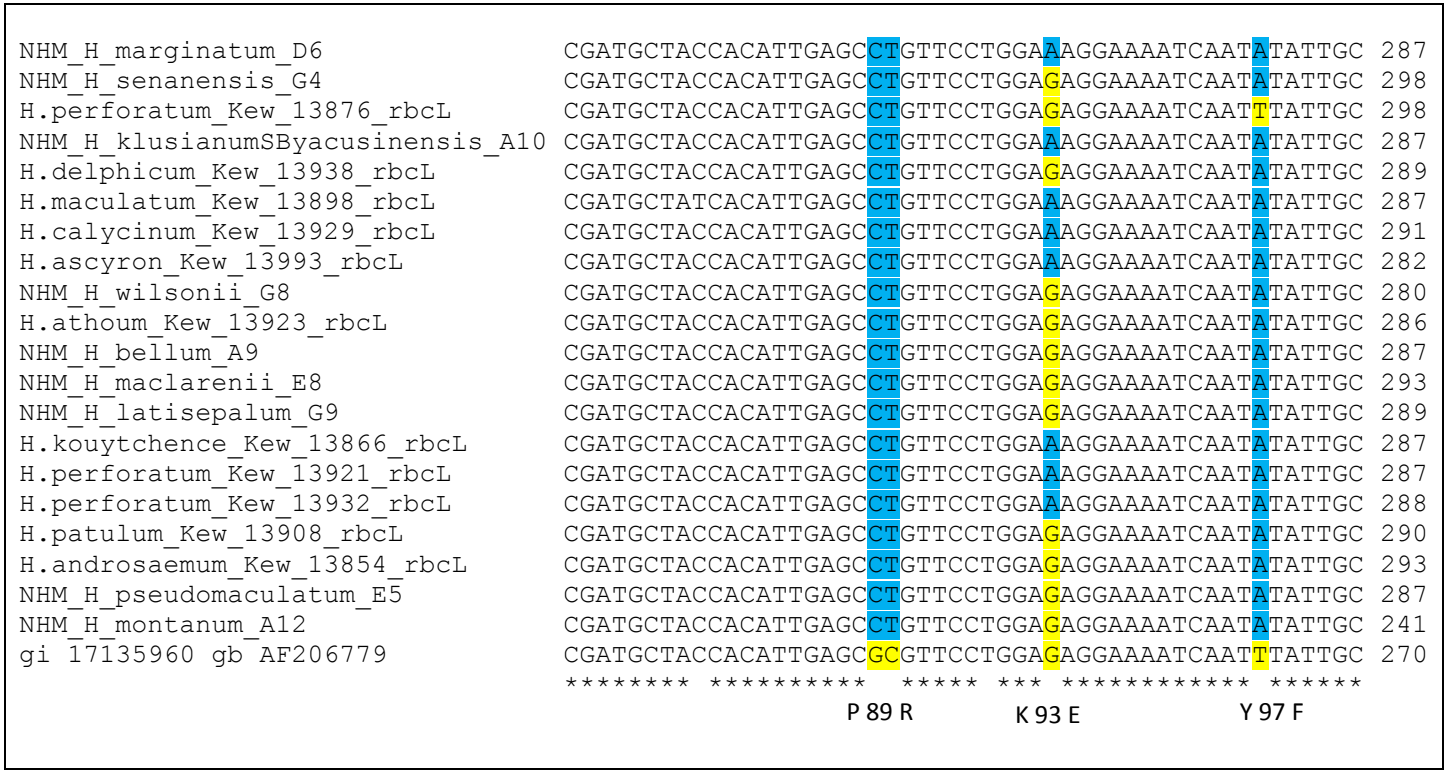


Figure 5.16 Section of nucleotide multiple alignment with four substitutions indicated which cause amino acid alterations.

AA 97; Y for F is the amino acid substitution indicated in Figure 5.15, caused by substitution of an A for a T nucleotide. Other base differences in this alignment are not highlighted as they do not cause AA substitutions.

Each individual nucleotide substitution which caused an alteration to the resultant AA sequences was investigated. An example of this is shown for one sample in Table 5.4 (full results are shown in Appendix section 8.1.4). The substituted AA was identified by its number in the protein alignment, and the AA it was substituted for. This enabled a judgement as to the likelihood of the AA change. This was aided by a multiple alignment of all the published *Clusiaceae* family *rbcL* protein sequences (Figure 5.17, shown in full in Additional Appendix). If the change was already observed within the family of sequences it was deemed more likely than those which were not. In addition, the nucleotide causing the change was identified and its base number in that sample's sequence determined. This base position was isolated in the consensus sequence in order to ascertain whether a conflict had occurred in calling that base and whether the resolution of that conflict had resulted in a correct base call. Conflicts occur when the sequences contributing to a consensus sequence have different base calls at one position. In most sequence analysis software, these conflicts are resolved based on 'vote' with each read having one vote. When two reads have been used assemble a contig, this vote inevitably results in a draw so the individual electropherogram traces must be assessed to call the base. Examples of this are shown in Table 5.5. Six of the nine substitutions in the sample shown in Table 5.5 were the result of vote resolved conflicts, with the majority of these being between A and G. In all of these instances the software called an A, which appears to be a tendency of the software towards A. On comparison of the traces, all of these base calls were reversed. All of the DNA sequences produced were corrected based on this type of analysis, as shown for one sample Table 5.4.

The adjusted sequences were then aligned (Figure 5.18), a high consensus was seen between all samples with 30 positions in total where different bases are seen throughout the alignment. Due to this high conservation, the few polymorphic sites available were assessed as to their ability to group species. A polymorphism which is present in several of the studied samples creates two distinct groups, these groups could form the beginning of a system similar to that of barcode traffic lights described by Chase et al. (Chase et al., 2005). This system used several barcodes, acting at hierarchical points, to result in a species level identification. For instance, one region may resolve a sample to an order, based on this a second region is chosen to resolve further and this is followed until a species level identification is established. The polymorphic sites could be used in a similar way based upon the nucleotide at each of a selection of positions. This type of analysis was conducted on the alignment, resulting in the seven positions of interest shown in Table 5.6.

Table 5.4 Example of the *rbcL* amino acid sequence analysis.

The AA substitution is characterised, the bases responsible identified, the electropherogram traces considered and the frequency of the difference, the call is then either accepted and no change made to the sequence, or it is rejected and the sequence is altered.

Source	Ref.	AA number	Different AA	Sequence difference	Trace Reading	Decision	Other sequences?	Action?
Kew	perf 921	10	S for G	19 - A	f = trim, r = A	A	Pub G = G, others A = S	Accept
Kew	perf 921	19	V for E	47 - T	f = A (trimmed), r = T	A	All others A = E	Reject
Kew	perf 921	89	P for R	257 - C, 258 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	Accept
Kew	perf 921	93	K for E	268 - A for G	f = A/G, r = G	G	8 K, rest E, at EE	Reject
Kew	perf 921	97	Y for F	281 - A for T	f = A, r = T	A	pub and 876 = F, rest = Y	Acept
Kew	perf 921	106	N for D	307 - A for G	f = A/G, r = G	G	pub and 3 = D, rest N	Reject
Kew	perf 921	110	K for E	319 - A	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Reject
Kew	perf 921	159	K for R	468 - A for G	f = A/G, r = G	G	6 = K, rest (inc pub) +R	Reject
Kew	perf 921	160	N for D	469 - A	f = A/G, r = G	Trim	5 (inc pub) = D, 1 = K, rest = N	Reject

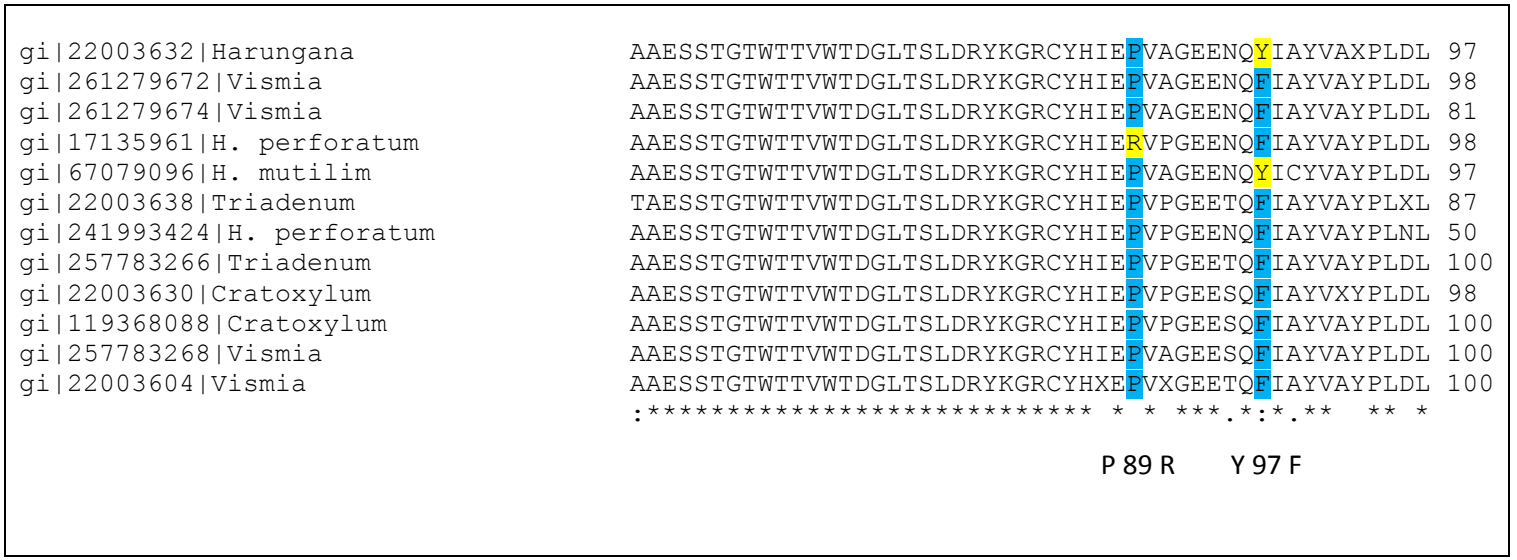
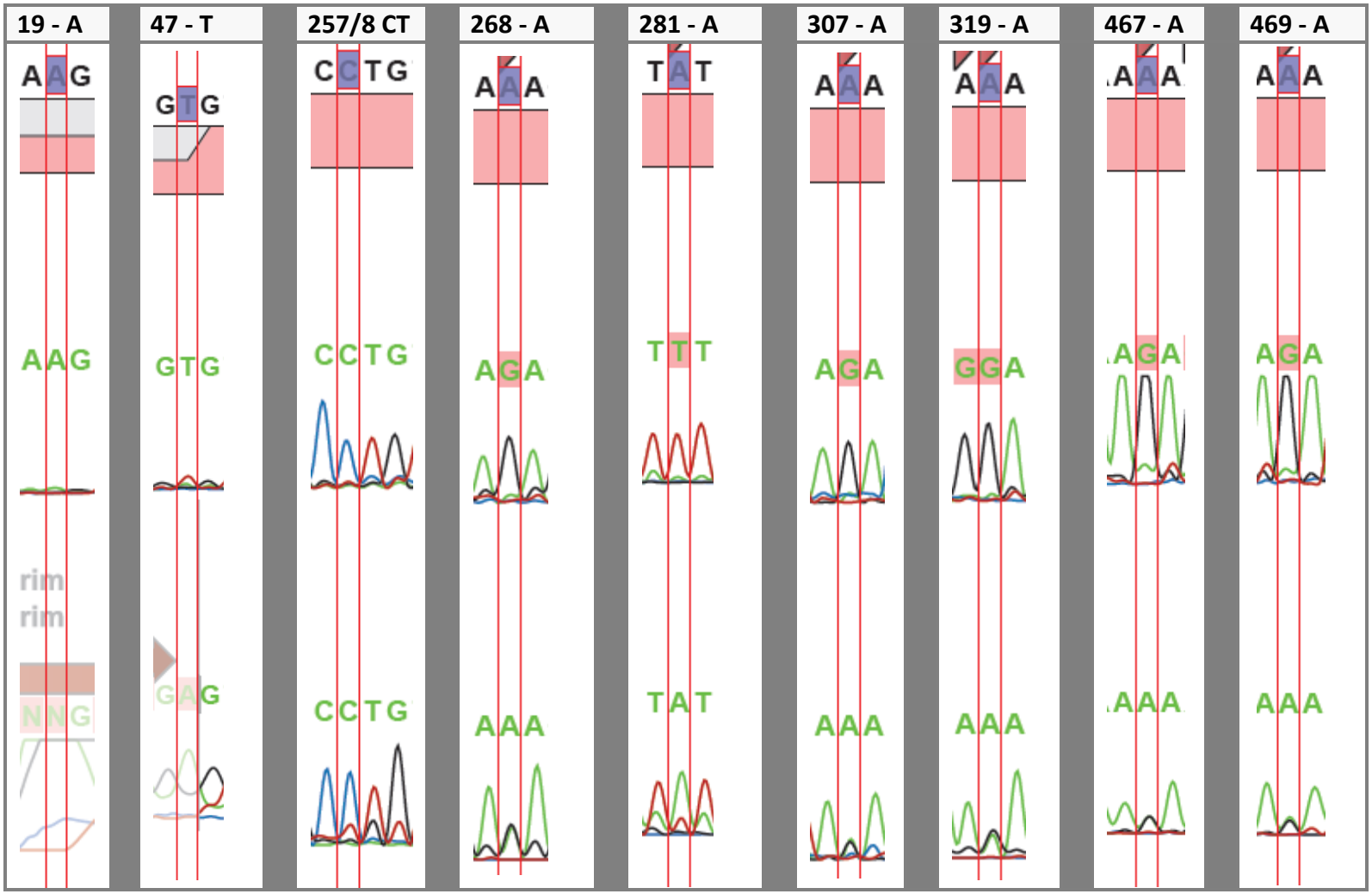


Figure 5.17 Section of the AA sequence alignment of published *Clusiaceae rbcL* sequences.

Two AA substitutions are indicated in blue and yellow, at position 89 the majority of published sequences have P – Proline, and one of the published *H. perforatum* sequences has R – Arginine. All the samples sequenced coded for P, and this was accepted as the correct call. At position 97 the consensus AA is F – Phenylalanine with two sequences coding for Y – Tyrosine. All but one of the samples sequenced in this study code for Y and this was accepted due to its dominance in the data and appearance in published sequences.

Table 5.5 Individual sequence differences causing AA alterations in the Kew *H. perforatum* 921 sample. The electropherogram trace for each base difference described in Table 5.4 is shown, enabling a second check of the base calling.



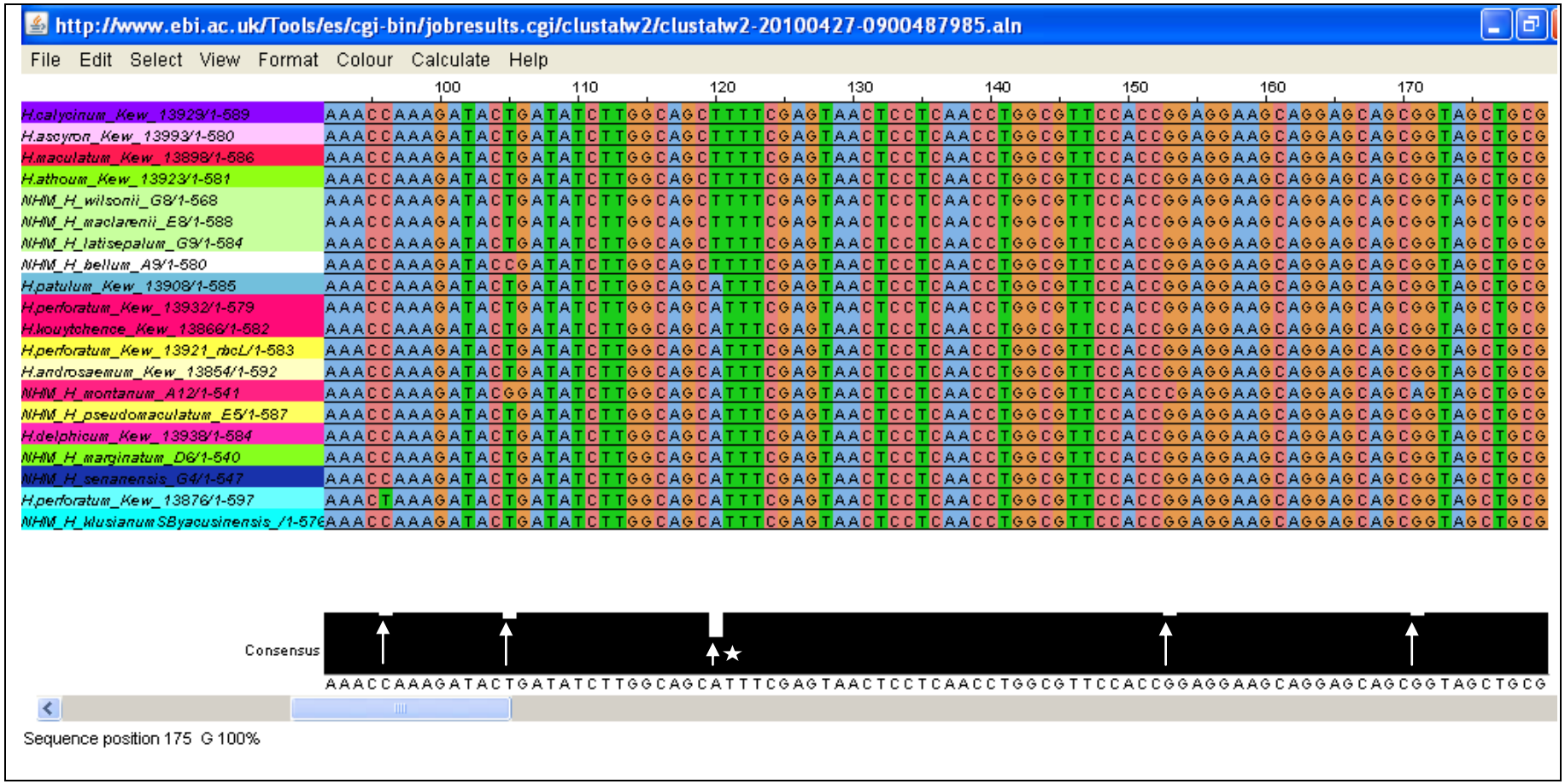


Figure 5.18 A section of the multiple alignment of the corrected *rbcL* sequences.

Five positions are shown with nucleotide variations as indicated by white arrows on the consensus sequence bar; an A/T substitution is indicated with a smaller arrow with a star which splits the sequences into two groups, the other four substitutions are confined to one or two sequences. The sites which cause a definite and large split between the sequences are considered to be the most useful polymorphic sites.

Table 5.6 Polymorphic nucleotide positions within the *rbcL* region for 19 *Hypericum* species.

The nucleotide at each polymorphic position creates a pattern, though these are not specific to each species. The samples fall into three haplogroups, indicated.

Hypericum species	Sample Source	Sample Reference	Polymorphism position						
			198	390	108	396	510	414	
maclarenii	NHM	E8	C	A	T	G	T	T	Haplogroup 1
latisepalum	NHM	G9	C	A	T	G	T	T	
athoum	Kew	13923	C	A	T	G	T	T	
wilsonii	NHM	G8	C	A	T	G	T	T	
maculatum	Kew	13898	C	A	T	G	T	T	
ascyron	Kew	13993	C	A	T	G	T	T	
bellum	NHM	A9	C	A	T	G	T	T	
calycinum	Kew	13929	C	A	T	G	T	T	
androsaemum	Kew	13854	C	A	A	A	A	C	Haplogroup 2
kouytchense	Kew	13866	C	A	A	A	A	T	
patulum	Kew	13908	C	A	A	A	A	T	
perforatum	Kew	13932	C	A	A	A	A	T	
klusianum sb. yacusinensis	NHM	A10	T	C	A	A	A	T	Haplogroup 3
perforatum	Kew	13876	T	C	A	A	A	T	
delphicum	Kew	13938	T	C	A	A	A	C	
pseudomaculatum	NHM	E5	T	C	A	A	A	C	
senanensis	NHM	G4	T	C	A	A	A	C	
marginatum	NHM	D6	C	C	A	A	A	T	
montanum	NHM	A12	T	C	A	G	A	T	

The data shown in Table 5.6 show that while these Single Nucleotide Polymorphisms (SNPs) provide genotype patterns for the different *Hypericum* species, they are not capable of resolving to the species level in all cases. While *H. androsaemum*, *H. marginatum* and *H. montanum* have specific patterns within the sample group, the vast majority of species do not.

They do fall into three main haplogroups, as indicated (Table 5.6), which could potentially be used to place unknown samples. For example, a sample sold as *H. maclarenii*, a Haplogroup 1 species, could be sequenced and should the banding pattern match haplogroup 2 or 3 it could be surmised that the sample was not *H. maclarenii*, although which particular species it is would remain unknown.

Interestingly, another polymorphism follows the SNP distribution of position 390 exactly, but it was removed based on unsound basecalling, position 268 A for G shown in Table 5.4. This brings in to question whether the unsound basecalling is just that, or in fact an indication of sequence differences present within the sample. The electropherogram traces would suggest not, however the distribution matching exactly to another polymorphism in extremely unlikely to have occurred by chance alone.

Also of particular interest are the two *H. perforatum* samples, one of which falls into haplogroup 2 (sample 13932) and the other into haplogroup 3 (sample 13876). This would suggest that despite high conservation of the *rbcL* region in general across the sampled *Hypericum* species, variation between members of the same species still occurs at a similar rate to variation between species. This is further discussed in section 5.3.5.5.

5.3.2.2 Lithuanian Samples

The *rbcL* region in all of the Lithuanian samples was extremely highly conserved, similarly to the Kew and NHM sample sets. The polymorphic sites identified in Table 5.6 were investigated along with one other position that appeared to be polymorphic within the Lithuanian samples. This analysis is shown in Table 5.7 along with the patterns for three of the Kew DNABank samples, two *H. perforatum* and one *H. maculatum*, for reference. Four main haplogroups of patterning are indicated on Table 5.7. Haplogroup 1 shares the first three chosen SNPs with Haplogroup 1 from Table 5.6; it contains two Kew samples, *H. perforatum* 13932 and *H. maculatum*, and one sample from each species from the Lithuanian set. Given a larger sample, this first group would be expected to form the bases of at least three independent pattern based groups. Haplogroups 3a, 3b and 3c vary only in the last two chosen SNPs; the first six positions all displaying the TCAAA pattern shown in Haplogroup 3 from Table 5.6. This group contains a Kew *H. perforatum* (13876) and *H. delphicum*, and the NHM *H. pseudomaculatum*. None of the groups are species specific, and the ratios of *H. perforatum* to *H. maculatum* samples in each group vary: Haplogroup 1 is 1:1, Haplogroup 3a is 1:2, Haplogroup 3b is 1:2

and Haplogroup 3c is 18:7. This again indicates that although the SNPs can be used to inform, they cannot identify to the species level.

Table 5.7 Polymorphic nucleotide positions in the *rbcL* region for *H. perforatum* and *H. maculatum* samples from the Lithuanian set and representative Kew samples

H. maculatum / *H. perforatum*	Sample Number	Polymorphism Position							
		198	390	108	396	510	414	281	
maculatum	Kew 13898	C	A	T	G	T	C	A	Haplogroup 1
maculatum	6	C	A	T	G	T	C	A	
perforatum	2	C	A	A	G	T	C	A	
perforatum	Kew 13932	C	A	A	A	A	C	A	Haplogroup 2
maculatum	7	T	C	A	A	A	C	A	Haplogroup 3a
maculatum	1	T	C	A	A	A	C	A	
maculatum	14	T	C	A	A	A	C	A	
maculatum	16	T	C	A	A	A	C	A	
maculatum	9	T	C	A	A	A	C	A	
maculatum	5	T	C	A	A	A	C	A	
perforatum	3	T	C	A	A	A	C	A	
perforatum	11	T	C	A	A	A	C	A	
perforatum	8	T	C	A	A	A	C	A	
perforatum	22	T	C	A	A	A	C	T	Haplogroup 3b
maculatum	2	T	C	A	A	A	C	T	
maculatum	10	T	C	A	A	A	C	T	
maculatum	3	T	C	A	A	A	T	T	Haplogroup 3c
maculatum	11	T	C	A	A	A	T	T	
maculatum	4	T	C	A	A	A	T	T	
maculatum	13	T	C	A	A	A	T	T	
maculatum	8	T	C	A	A	A	T	T	
maculatum	12	T	C	A	A	A	T	T	
maculatum	15	T	C	A	A	A	T	T	
perforatum	1	T	C	A	A	A	T	T	
perforatum	15	T	C	A	A	A	T	T	
perforatum	10	T	C	A	A	A	T	T	
perforatum	5	T	C	A	A	A	T	T	
perforatum	18	T	C	A	A	A	T	T	
perforatum	14	T	C	A	A	A	T	T	
perforatum	4	T	C	A	A	A	T	T	
perforatum	13	T	C	A	A	A	T	T	
perforatum	9	T	C	A	A	A	T	T	
perforatum	21	T	C	A	A	A	T	T	
perforatum	16	T	C	A	A	A	T	T	
perforatum	6	T	C	A	A	A	T	T	
perforatum	17	T	C	A	A	A	T	T	
perforatum	20	T	C	A	A	A	T	T	
perforatum	12	T	C	A	A	A	T	T	
perforatum	19	T	C	A	A	A	T	T	
perforatum	7	T	C	A	A	A	T	T	
perforatum	Kew 13876	T	C	A	A	A	T	T	

5.3.3 *matK* region

The second designated barcode, the *matK* region, was very difficult to amplify in all the *Hypericum* samples studied. All available published primer sets were tested (Table 5.8), in all possible combinations. Different protocols were also tested for each, with cycle numbers ranging up to 40, and the addition of the adjuvant DMSO at various concentrations up to 10%.

Table 5.8 Table of published primers trialled with *Hypericum* species and their sources.

All Kew primer sequences are freely available on-line at http://www.kew.org/barcoding/. Full sequences are shown in section 5.2.2.1.

Primer	Direction	Source
2.1	F	Kew Barcoding Phase 1
2.1a	F	Kew Barcoding Phase 1
X	F	Kew Barcoding Phase 2
5	R	Kew Barcoding Phase 1
3.2	R	Kew Barcoding Phase 1
390F	F	Cuenoud et al. 2002
1326R	R	Cuenoud et al. 2002
3F_KIM f	R	Lahaye et al. 2008
1R_KIM r	F	Lahaye et al. 2008

The most successful attempt used the primers 390F and 1326R (Cuenoud et al., 2002), and a double PCR procedure using an initial amplification as the target for a further PCR reaction, beginning the reaction with very low stringency parameters. This results in non-specific amplification, with multiple banding as indicated in Figure 5.19. The band of the expected approximate base pair size was excised from the gel, cleaned and used as the target for cycle sequencing.

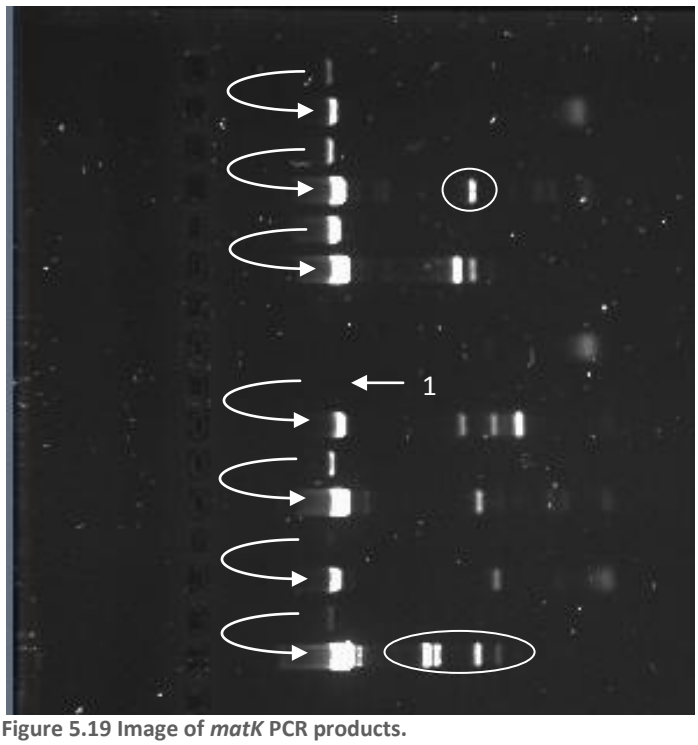


Figure 5.19 Image of *matK* PCR products.
In each case, the first product acts as the template for the second reaction, indicated by arrows. The second reaction increased the target band in all cases, in the example indicated by a number 1 the first reaction produced no visible band. The second reaction in all cases resulted in multiple bands, some are indicated by circles.

Initial attempts at cycle sequencing using the amplification primers were unsuccessful. As of August 2009, one published *matK* sequence was available of GenBank for *H. perforatum*, Accession no. DQ168438.1. This was used as a template for the design of sequencing primers to be used with the amplified products. The primers were designed to anneal at either end of the sequence, base number 1 and 949, and in the centre, base number 446. This was to enable sufficient overlap of sequence reads as the sequence was known to be difficult to sequence.

Table 5.9 Sequencing primers designed for matK region of Hypericum.

Annealing position in the sequence is indicated in the primer name.

Primer name	Direction	Primer Sequence
HPmK 446F	F	TTCAAACCCTTCGGTACTGG
HPmK 446R	R	CCAGTACCGAAGGGTTTGAA
HPmK 949R	R	TTGGTTGAACCCACACAAAA
HPmK 949F	F	TTTTGTGTGGGTTCAACCAA
HPmK 1F	F	ATGCGGGAAGAGGAATATCA

These primers were all tested, with and without DMSO, with similar results obtained for all primers. The sequence begins as usual, but after only approximately 50 bases of reading the sequence falls in quality dramatically. At this point base calls cannot be reliably made, and the amount of sequence data produced is insufficient to be meaningfully analysed (Figure 5.20 and Figure 5.21).

Further work on reliable primers for this region in *Hypericum* species is required to assess whether this barcode could effectively identify *Hypericum* species, but currently this is not an option.

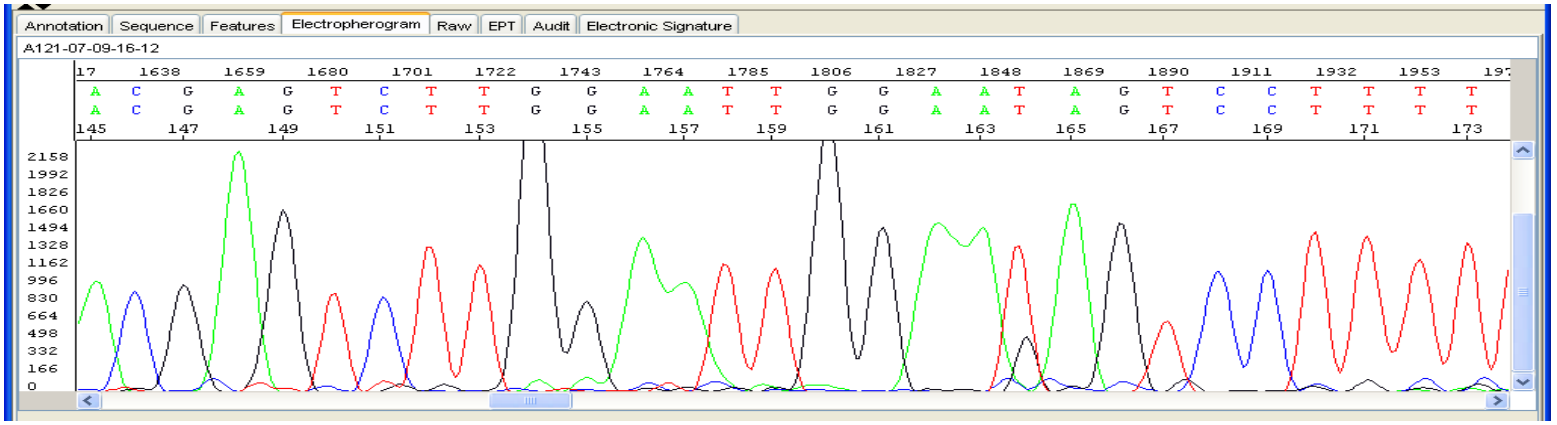


Figure 5.20 Section of electropherogram of *matK* cycle sequencing reaction products from base number 145 to 173.

After an initial phase of unreliable sequence, always seen in capillary sequencing, for approximately 50 bases a reasonable quality of sequence data is produced with very low 'background noise' and very few secondary peaks.

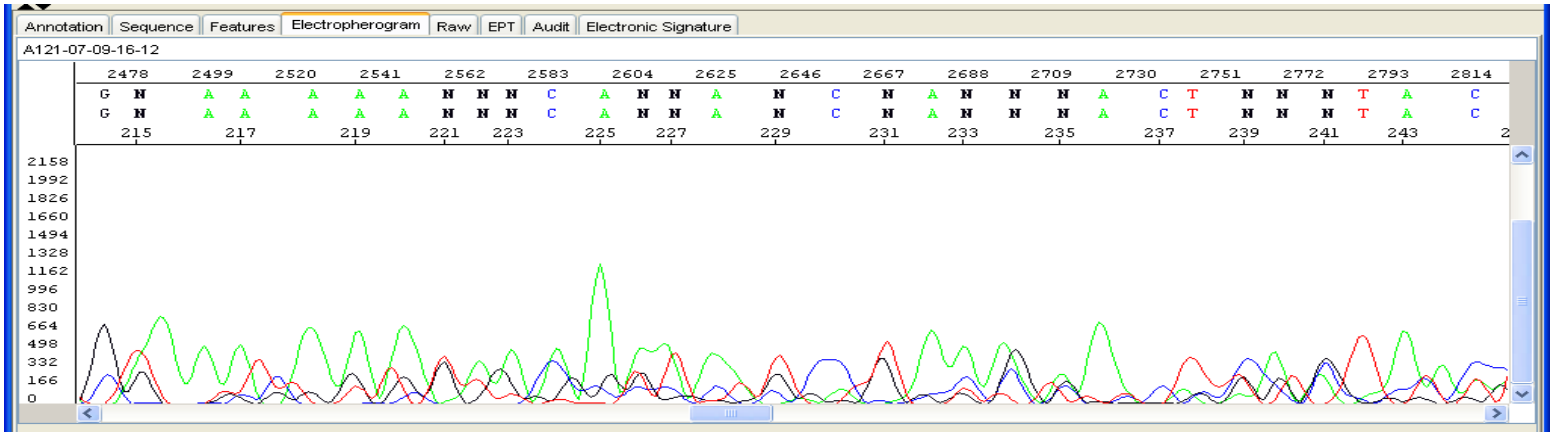


Figure 5.21 Section of electropherogram of *matK* cycle sequencing reaction products from base number 215 to 241.

This section is just 40 bases on from that shown in Figure 5.20, the sequence data has degenerated severely with no reliable base calls, and background noise is inseparable from target sequence and analysis impossible.

5.3.4 *trnH-psbA* spacer region

Amplification of the *trnH-psbA* region was not straight forward and required intensive optimisation; in order to universally amplify all DNA samples a 'touchdown' PCR cycling programme was used. Touchdown PCR begins with an initial annealing temperature higher than the *Tm* of the primers which results in an extremely high stringency reaction. The annealing temperature is then reduced by 1°C per cycle allowing exponential amplification of the initial products. Once a set annealing temperature is reached, cycling continues as usual. This method greatly increased amplification of this region, as shown in Figure 5.22.

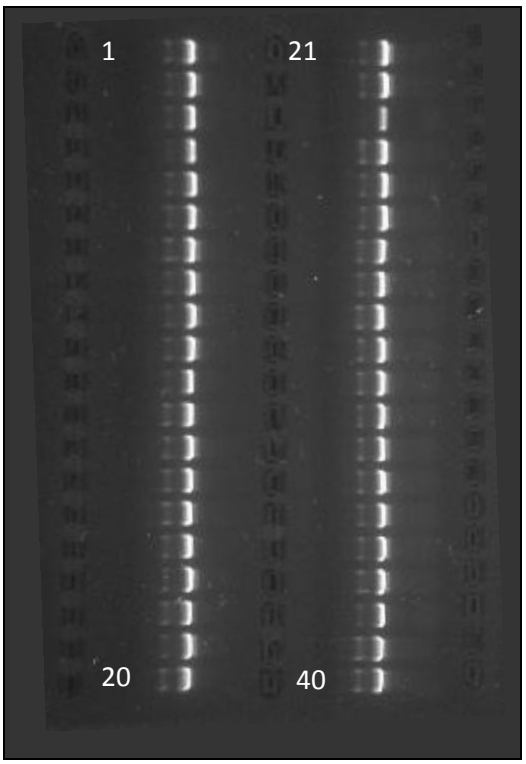


Figure 5.22 Image of gel with *trnH-psbA* amplification products for all 38 Lithuanian samples and two others in lanes 39 and 40.

The *trnH-psbA* intergenic spacer region has been reported as a difficult region to sequence due to mononucleotide repeats and small inversion and duplication events (Fazekas et al., 2008). The quality of the data obtained in this investigation was similar to that from the nrITS region; samples are shown in Figure 5.23 and Figure 5.24. This is in agreement with CBOLs findings that this region produces lower than optimum quality sequence data (Hollingsworth et al., 2009, CBOL et al., 2009).

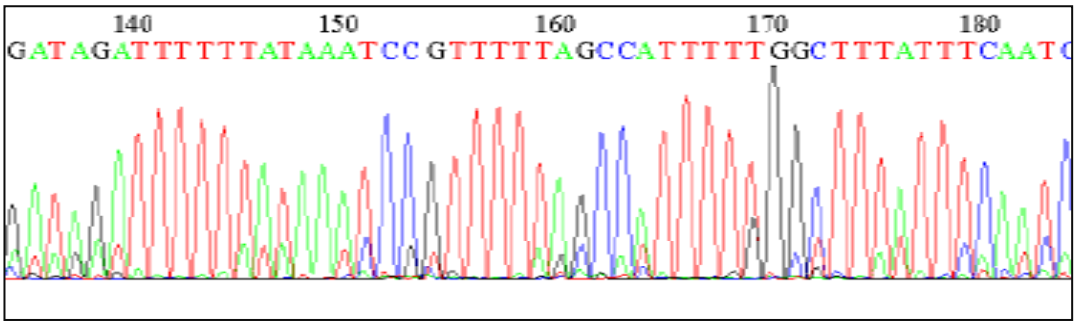


Figure 5.23 Section of the electropherogram from an external forward primer *trnH-psbA* cycle sequencing reaction Macrogen image output.

Secondary peaks are visible at many positions and the overall fluorescence levels vary dramatically.

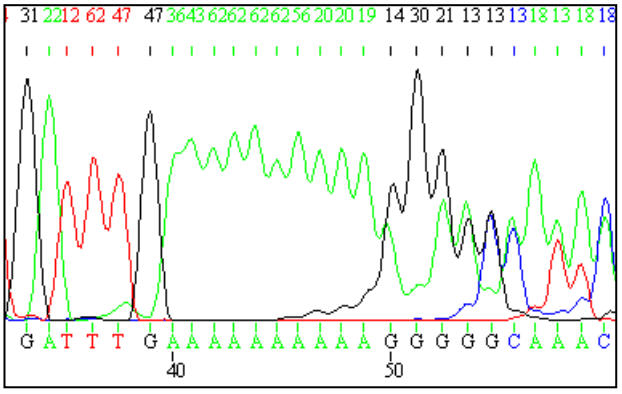


Figure 5.24 Section of the electropherogram from a reverse *trnH-psbA* cycle sequencing reaction Macrogen PDF output.

An 'A' mononucleotide repeat is shown followed by a region with very difficult base calls due to multiple peaks.

The electropherogram traces show that at several base positions a second smaller peak is visible; this is the reason the data are described as lower than optimal quality. However, these are corrected by the use of multiple reads to create a final consensus sequence. These were assembled from two forward and one reverse read for this region. The regions sequenced by each of these primers are shown in Figure 5.25, with each overlapping over a significant proportion.

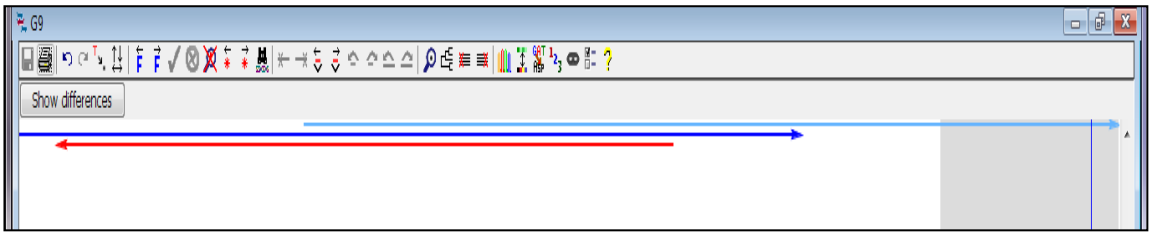


Figure 5.25 A section of the CodonCode Aligner output for *trnH-psbA* region.
The different reading regions from sequencing primers are shown; HTPSF dark blue, HTPSF2 light blue and HTPSR red.

Incorporating all of these methods, analysis quality data was produced for 91% of the Kew samples, 87% of the NHM set and 100% of the Lithuanian sample set.

5.3.4.1 Vouchered samples

The sequences from all 31 vouchered DNA samples were aligned along with the only published *Hypericum trnH-psbA* region available on GenBank as of 8/03/2010, accession number DQ006195.1, for the species *H. mutilum* (Figure 5.26, shown in full in Additional Appendix). The alignment of these regions is problematic with sequence lengths ranging from 595bp to 968bp, as has been described in previous studies (See section 5.3.3). Small inversion and duplication events cause large shifts in the alignment and are also not read efficiently by alignment software, Figure 5.26 shows examples of both of these events. Sequence differences between species are frequent but rarely species specific, and the nature of these differences as described above, is incompatible with many DNA based identification methods. DNA based techniques depend upon recognition of precise nucleotide sequences, so when a region is constructed of duplications, inversions and mononucleotide repeats all resolving power is lost as recognition sequences may appear multiple times and in different orientations.

A phylogenetic tree based on percent identity was produced from this multiple alignment, Figure 5.27. This groups the species into three main groups, the first of which is made up predominantly of species from the *Adenosepalum* section of the *Hypericum* genus. Groups 2 and 3 show a mixture of sections, most notably between the three *H. perforatum* samples included. *H. perforatum* 932 and 921 are both in sections 2, whereas *H. perforatum* 876 is in section 3. This separation matches that seen in the *rbcL* polymorphic markers, and suggests that plastid separation within the *H. perforatum* species is not complete.

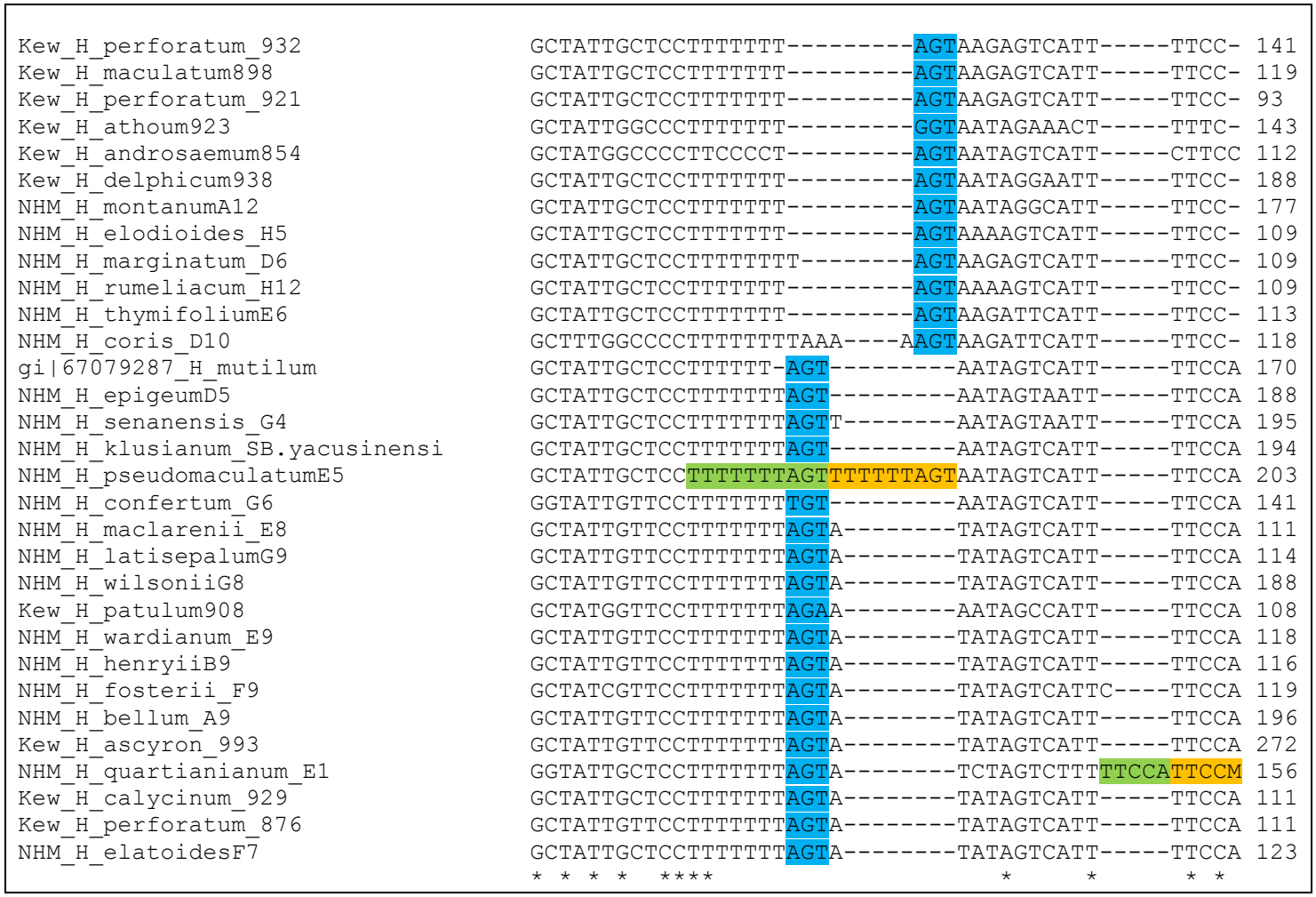


Figure 5.26 Section of the multiple alignment of the *trnH-psbA* sequences for the Kew and NHM sample sets.

The triplet highlighted in blue could have been more meaningfully aligned manually; demonstrating one of the problems with this region, the difficulty of alignment. Shown in orange and green are two probable duplication events, in each case the green appearing to be a copy of the orange although it is impossible to know which copy is the original.

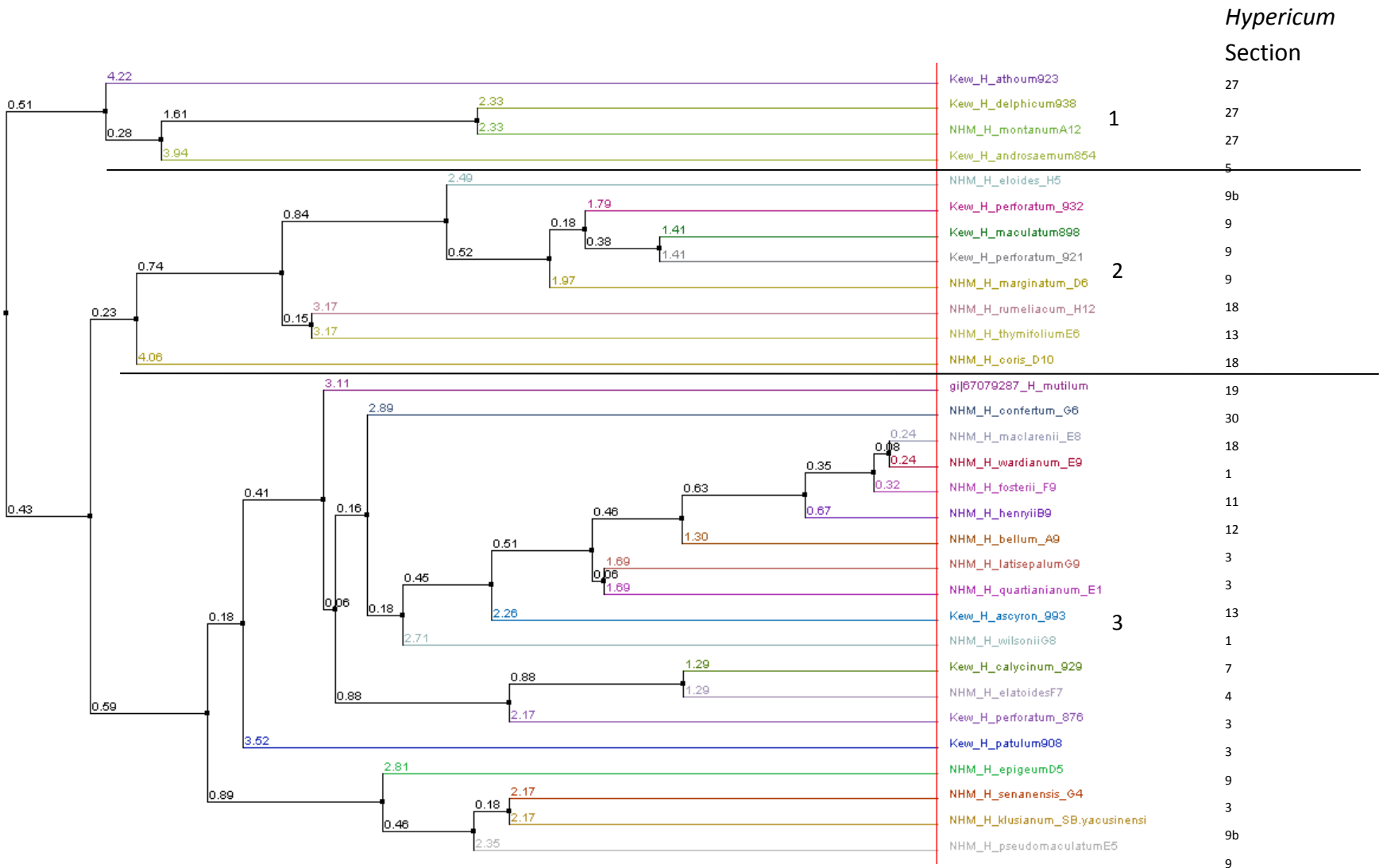


Figure 5.27 Average distance tree created using Percent ID for the Kew and NHM sample sets over the *trnH-psbA* region.

The samples separate into three groups, indicated. The Section of each species within the *Hypericum* genus is shown on the right.

5.3.4.2 Lithuanian Samples

The Lithuanian samples *trnH-psbA* sequences were aligned and regions characteristic of the *H. maculatum* samples were apparent, an example being shown in Figure 5.28 (Full alignment shown in Additional Appendix). Similarly, *H. perforatum* type regions were identified. In the example shown in Figure 5.29 this is a sequence gap caused by a deletion rather than a species specific sequence. These sequence differences enabled this region to be used to separate the Lithuanian samples into two groups, one made up entirely of *H. maculatum* samples and the other predominantly *H. perforatum*, although the split is not perfect (Figure 5.30). This is in contrast to the results from the three vouchered *H. perforatum* samples discussed in section 5.3.4.1, which did not group together based on this region. Within this group of samples, less sequence variation is found than within the three vouchered samples.

Of the regions sequenced and analysed for these sample sets, the only one capable of any type of useful resolution of the Lithuanian samples has been *trnH-psbA*.

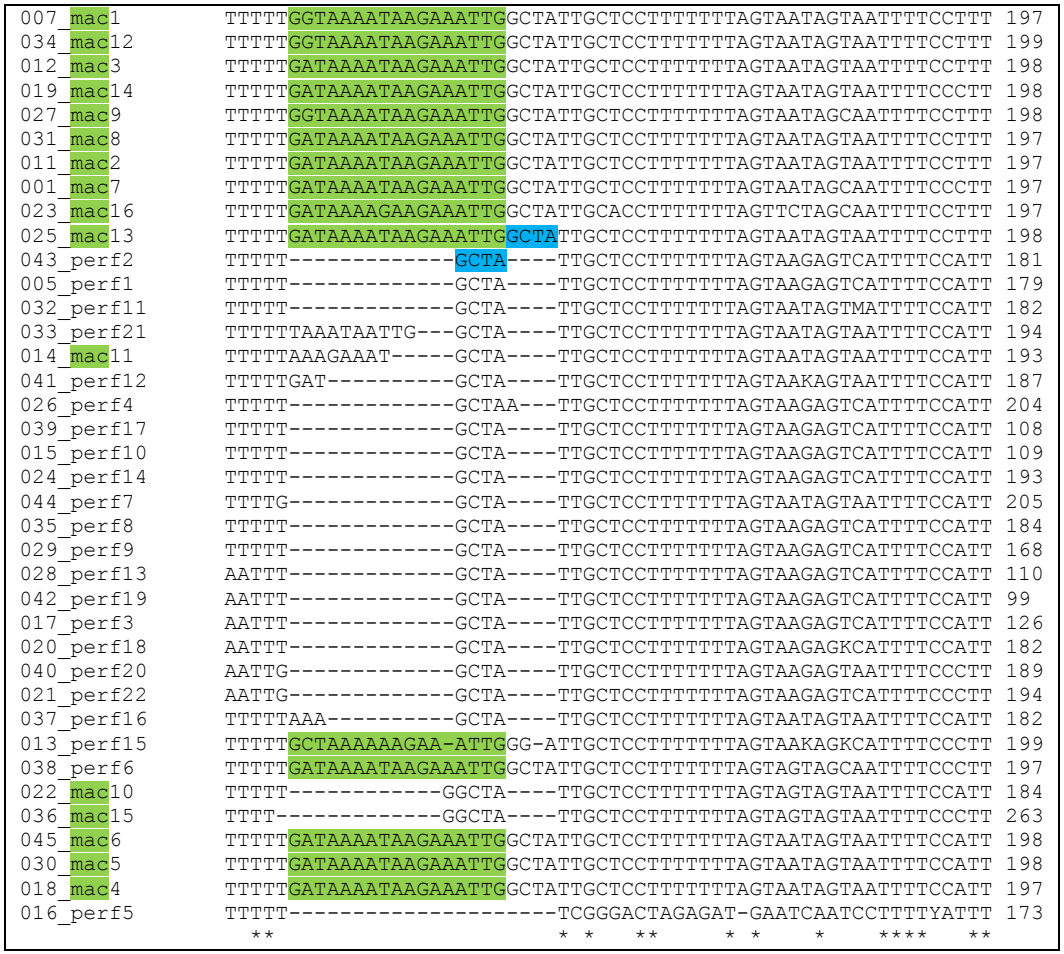


Figure 5.28 Section of the multiple alignment for the *trnH-psbA* region of the Lithuanian samples showing a feature of the *H. maculatum* sequences in green.

This feature is found in two *H. perforatum* sequences and is absent from three *H. maculatum* sequences but does aid separation of the samples. Alignment difficulties are still evident, as indicated in blue.

Figure 5.29 Section of the multiple alignment for the *trnH-psbA* region of the Lithuanian samples showing a feature of the *H. perforatum* sequences.

This is shown in green, an area conspicuous by its absence. This feature is found in all but one *H. perforatum* sequence and in just one *H. maculatum* sequence.

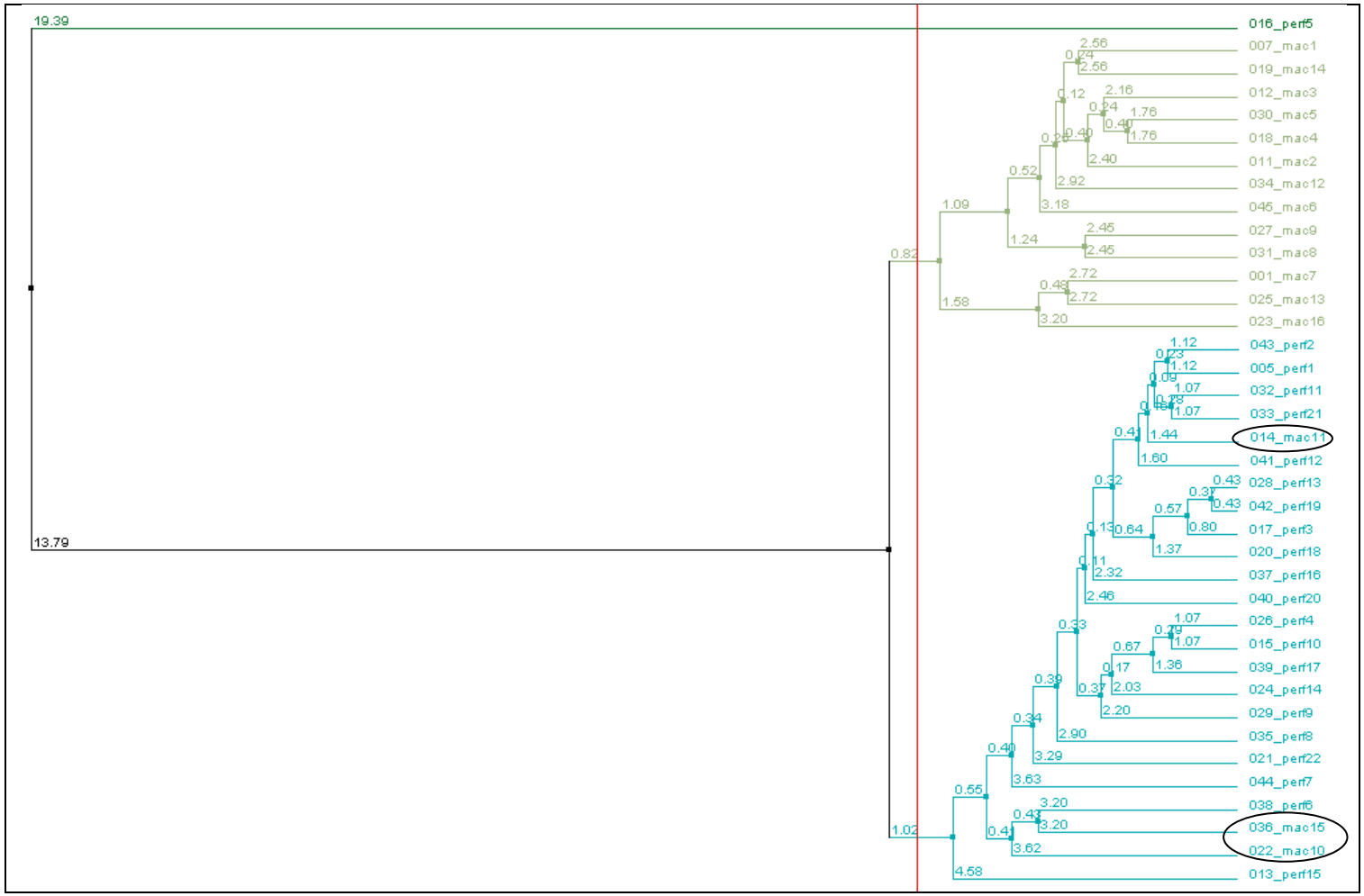


Figure 5.30 Average distance tree using Percent ID for the Lithuanian sample set over the *trnH-psbA* region.

This region separates the samples into two main groups with one outlying sample. One group is predominantly *H. perforatum* and the other *H. maculatum* making this the only region tested capable of separating these samples; albeit not perfectly as the samples mac10, mac15 and mac11 fall into the *H. perforatum* group.

5.3.5 Conclusions

5.3.5.1 nrITS Phylogeny of Hypericum

The results of the nrITS sequence analysis for the NHM samples showed broad agreement with both the Clade classification of Crockett et al. 2004 and the morphological proposal of Dr NK Robson. They also show that the split between these Clades is not as clear as had previously been reported, with some sections represented in both Clade A and Clade B. The species studied in this research covered fourteen of the thirty-six sections within the genus *Hypericum*. The inclusion of samples from all sections of the genus may indicate that Clades A and B should be subdivided or reclassified.

Further work including all of the sections is also likely to reveal that Clades A, B and C do not represent the entire genus. For example, the sample from the Old World *Hypericum* genus section *Androsaemum* is an outlier to the samples within Clades A and B, though would be unlikely to fall within Clade C as this is represented by only one New World section of the genus. This indicates that at least one more Clade would be identified with the inclusion of further accessions from this section.

5.3.5.2 nrITS sequence conservation between H. maculatum and H. perforatum

The nrITS region of the Lithuanian sample set showed extremely high sequence conservation and revealed this as the cause for the apparently "false" positive result of the microcode PCR assay. As the microcode assay has now been shown to identify *H. perforatum* and some *H. maculatum* samples, in order to ascertain whether this is due to geographical differences in the DNA sequence of *H. maculatum* further investigation is needed. This should be conducted with particular emphasis on the areas in which *H. perforatum* is cultivated, and whether local *H. maculatum* species are likely to mistakenly become involved in the production process.

5.3.5.3 Universal Amplification

Of the four regions tested, the highest amplification success rate was achieved with the *trnH-psbA* region, as sequence quality data was produced for 93% of all samples. However, this was only accomplished after intensive optimisation of the amplification protocols used. The *rbcL* region was the easiest to amplify and produced analysis quality data for 100% of the Kew and Lithuanian samples. In contrast, the NHM sample set produced sufficient quality for just 39% of the samples, resulting in *trnH-psbA* having the highest amplification success rate. This is likely to be due to the order of investigation rather than the primers and protocols used, as the NHM sample set was extremely low in quantity and the *rbcL* region was the last to be

sequenced. Both the *rbcL* and *trnH-psbA* regions can now be easily and efficiently amplified and sequenced in *Hypericum* species based on the work described in this chapter.

5.3.5.4 DNA sequence variation and species identification

Reliable sequence data was produced for the Lithuanian set of *H. perforatum* and *H. maculatum* samples for three regions, of these the most capable of separating the samples by species was *trnH-psbA*. Albeit that this resolution was not perfect, as three *H. maculatum* samples were grouped within the *H. perforatum* section. The separation of the samples into species groups required the entire *trnH-psbA* DNA sequence, and a phylogenetic approach to the data analysis producing a tree based on percent identity of sequences using neighbour joining. This is similar to the identification procedure foreseen with DNA barcoding, in which sequences from samples of unknown species are searched against the BOLD database, and identification made on the grounds of the closest match. Unfortunately, the complications of aligning the *trnH-psbA* region are likely to prevent the successful application of this system as the number of sequences for comparison is increased, as described by CBOL when deciding not to select this region as a plant barcode (Executive Committee, 2009).

The *rbcL* and nrITS regions showed similar degrees of conservation within the Lithuanian sample set, with just a few base differences across the entire range. Although this sequence conservation prevents the discrimination of the two very closely related species *H. perforatum* and *H. maculatum*, it does show that both of these regions are stable within their respective genomes. This would enable the anchoring of DNA-based techniques within these regions, with more highly variable and less stable regions acting as secondary markers to define samples to a lower taxonomic level.

None of the three regions sequenced and analysed were completely suitable as DNA barcodes. However, the nrITS region sequence variation between species was generally at the level required for barcode analysis, and the within species conservation extremely high. Only the inability to differentiate between the Lithuanian *H. perforatum* and *H. maculatum* samples prevents the whole-hearted recommendation of nrITS within *Hypericum* species.

The use of the nrITS is not recommended in barcoding due to large sequence differences causing alignment difficulties in some plants (Sass et al., 2007), and a complicated evolutionary pattern (Chase et al., 2007). The complex nature of the molecular evolution of the nrITS is due to the fact that multiple copies are present within the nuclear genome. They are part of the

nucleolus organising regions (NOR) which are found widely distributed in eukaryotic genomes (Alvarez and Wendel, 2003). The presence of many copies of the region within the genome can cause problems, as sequence differences between the copies would only be resolved by cloning of all the copies present and individually sequencing them (Feliner and Rossello, 2007). However, the nrITS region is known to undergo concerted evolution, a process by which the copies of nrITS are homogenised by cross-over events (Feliner and Rossello, 2007, Alvarez and Wendel, 2003). The sequence difference problems occur only when this process is incomplete, which does not appear to be the case within *Hypericum*. Within this research, and the studies of Crockett et al. (2004) and Park and Kim (2004), the nrITS region was stable and easily sequenced. The high degree of conservation between the sequences produced in this research and those published also suggests that the nrITS is a suitable region for the study of the *Hypericum* genus.

The three regions sequenced and analysed do all have different properties which are of potential use as platforms for DNA based identification assay design. The most suitable region was the nrITS, the utility of which has been shown in previous chapters. The *rbcL* data present the possibility of an entirely new method of identification based on SNP detection and mapping, which could be developed using new techniques such as SNaPshot™ from Applied Biosystems. This technique uses fluorescently labelled ddNTPs in PCR reactions to detect SNPs based on separation of the resultant fragments by Capillary Electrophoresis, this enables the typing of up to ten SNPs within one reaction (Applied Biosystems, 2000).

The *trnH-psbA* region in its entirety showed the greatest resolving power, and would be applicable to very similar samples. Further investigation may show that the nature of the *trnH-psbA* duplication and inversion events could also be utilised, similarly to SSRs (see section 1.3.6), if the region were stable enough within species.

5.3.5.5 Phylogenetic Inference from Plastid Markers

The phylogenetic analysis of the nrITS data largely agreed with the previously published Clades of Crockett et al. 2004, and in turn the morphology based resolution of the genus presented by Dr NK Robson. The inheritance of the chloroplast genome is generally accepted to be maternal (Chase and Fay, 2009), and certainly uniparental, as opposed to the nuclear genome which is biparentally inherited. Due to this, it is to be expected that the phylogenetic indications of plastid markers would differ from nuclear patterns.

However, it would be expected that the plastid regions of *rbcL* and *trnH-psbA* would show similar phylogenetic relationships to each other in these species, as they are both plastid regions. This is not the case, with species seeming to cluster into unrelated groups based on either region (Table 5.10). For example the *Hypericum* section *Ascyreia* clusters well based on the *trnH-psbA* region into group 3, but these species are assigned to two different *rbcL* groups.

There are some trends within the data; of note is the assignment of the *H. perforatum* sample 876 into a separate group to the other two *H. perforatum* samples, 921 and 932, based on both the *rbcL* and *trnH-psbA* regions. This is itself surprising, given the extremely high conservation seen in the *rbcL* region in general, and presents two conflicting indications. One the one hand the *rbcL* is highly conserved between species, implying a very low rate of molecular evolutionary, and on the other, the difference in sequence within one species implies that the sequence have altered since species divergence. The *trnH-psbA* region also presents inconsistencies, as despite being the only region capable of reliably grouping the Lithuanian samples by species, it also separates the three vouchered *H. perforatum* samples.

There are many possible reasons for these discrepancies; the first to be considered is always the possibility of sample mismanagement or errors in data handling. This is argued against in this case based on the continued use of samples in different assays throughout the research without any results indicating that samples were not as labelled or had been contaminated (See Chapters 2, 3, and 4). The preparation of samples for DNA sequencing was conducted both personally and by a supervised research assistant, both producing compatible results indicating that researcher bias or systematic errors are unlikely. DNA sequencing was conducted in-house and out sourced to a commercial company, producing congruent data. These data from different sources and using different sequencing primers for repeated DNA sequence reads for each region studied were assembled into contigs, with up to three sequence reads contributing to any one sequence. Had contamination or sample management errors occurred the assembly of contigs would have been impossible, as input sequences must be extremely similar for the software used to produce a consensus sequence. Large differences in sequence cause the assembly to fail, and no contig is produced. All of these factors indicate that the sequence data produced are accurate.

The next most parsimonious cause for the data discrepancies is the original misidentification of the vouchered samples, is *H. perforatum* 876 actually *H. perforatum*? The results of the microcode PCR test also indicate that the vouchered *H. maculatum* sequence differs from both

the Lithuanian *H. maculatum* samples and the published DNA sequence within the nrITS region, so could this also be caused by misidentification?

The only region capable of distinguishing the Lithuanian *H. perforatum* and *H. maculatum* samples was *trnH-psbA*, though this region separated the three *H. perforatum* samples when the entire sequence is used. The Lithuanian data alignment enabled the identification features within the sequence which could be grouped as *H. maculatum* or *H. perforatum* type, so do the vouchered samples align at these regions and which type are they? A multiple alignment was conducted to test this, a section of which is shown in Figure 5.31 with the *H. maculatum* and *H. perforatum* type regions highlighted. The alignment has shifted with the different input sequence, as compared to Figure 5.28 and Figure 5.29, causing the *H. maculatum* region to become partially aligned with the other sequences. Unexpectedly, the most similar sequence in this region to the Lithuanian *H. maculatum* is *H. perforatum* 876, the two sequences also align well accross the *H. perforatum* deletion. Equally, the *H. maculatum* 898 sequence aligns with the *H. perforatum* sequences in these regions.

It would be surprising, given the safeguards in place and expertise involved, if the identification of two of the vouchered DNA samples were from misidentified plants. This is also unlikely given the use of these samples throughout this research, during which they have consistently given the expected results based on the published sequences for each of these species. Further research into this should sequence the nrITS of these samples, and assess which species these samples are most similar to in both instances.

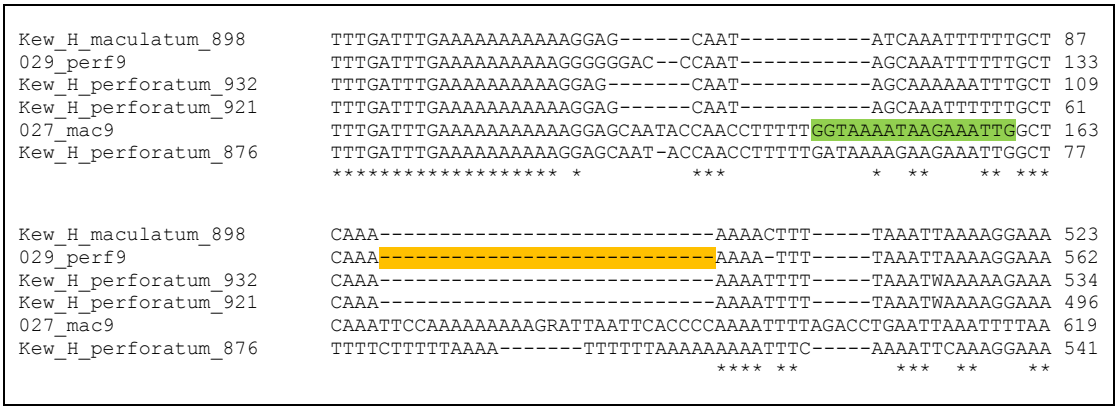


Figure 5.31 Section of the multiple alignment of the *trnH-psbA* region of the vouchered *H. perforatum* and *H. maculatum* samples with two representative Lithuanian samples.

The Lithuanian samples included were chosen as they exhibited the species indicative regions in Figure 5.28 and Figure 5.29. The *H. maculatum* type region is indicated in green, and the *H. perforatum* type region in orange.

Given the rejection of sample or data mismanagement and plant misidentification as possible causes for the discrepancies in the data, could the sequence differences and the resultant phylogenetic inferences from this be genuine?

The inheritance and segregation of the plastid genome has been reported not to follow the usual maternal inheritance pattern in some plant species. In some cases, paternal and even biparental inheritance has been reported (Hansen et al., 2007).

Similar data to those produced in this study were found to occur in the *Macaranga* species (Banfer, 2006), with six chloroplast haplotypes found in one species, and conversely the sharing of one haplotype between seven different species. Two possible explanations were examined to explain this; incomplete lineage sorting and hybridisation.

Incomplete lineage sorting results in the sharing of an ancient polymorphism between different species (Banfer, 2006). It occurs in species which have recently diverged and contain multiple different chloroplast types. Over time a consensus of chloroplast is reached by breeding within the species, but soon after divergence these different types of chloroplast can give similar results to those found in this research. Hybridisation events can cause similar findings when 'morphology capture' has occurred (Banfer, 2006). Pollination of the local flora by a new species over time results in this new chloroplast becoming the predominant chloroplast type within the area, resulting in a geographical distribution of chloroplast type rather than species based. This explanation best fits the results shown in this data set, as much less variation is seen between the Lithuanian samples compared to the vouchered samples.

Much further work is required to reveal the genealogy of the chloroplast within *Hypericum*, with many more regions sequenced and analysed with dedicated software. These data present an interesting finding, and a motivation for large scale research in this field. The prevalence of shared chloroplast genotypes should also be assessed in detail with particular reference to the DNA barcoding effort, as the two selected region are both within this genome.

Table 5.10 Assignment of Hypericum samples into groups based on the Hypericum genus sections, *rbcL* polymorphism patterning and *trnH-psbA* phylogenetic distance based tree.

Species/sample	*Hypericum* Section	Section number	*rbcL* Haplogroup	*trnH-psbA* group
H. quartianianum	*Campylosporus*	1	n/a	3
H. wardianum	*Campylopus*	1	n/a	3
H. maclarenii	*Campylosporus*	1	1	3
H. henryii,. Henryii	*Ascyreia*	3	n/a	3
H. elatoides	*Ascyreia*	3	n/a	3
H. monogynum	*Ascyreia*	3	n/a	n/a
H. bellum	*Ascyreia*	3	1	3
H. calycinum	*Ascyreia*	3	1	3
H. patulum	*Ascyreia*	3	2	3
H. kouytchense	*Ascyreia*	3	2	n/a
H. wilsonii	*Takasagoya*	4	1	3
H. androsaemum	*Androsaemum*	5	2	1
H. ascyron	*Roscyna*	7	1	3
H. maculatum	*Hypericum*	9	1	2
H. perforatum921	*Hypericum*	9	n/a	2
H. perforatum932	*Hypericum*	9	2	2
H. senanensis	*Hypericum*	9	3	3
H. perforatum876	*Hypericum*	9	3	3
H. pseudomaculatum	*Graveolentia*	9b	3	3
H. elodioides	*Graveolentia*	9b	n/a	2
H. epigeum	*Graveolentia*	9b	n/a	3
H. filicaule	*Monanthema*	9e	n/a	n/a
H. fosterii	*Origanifolia*	12	n/a	3
H. laxiflorum	*Origanifolia*	12	n/a	n/a
H. rumeliacum	*Drosocarpium*	13	n/a	2
H. latisepalum	*Drosocarpium*	13	1	3
H. klusianum, yacusinensis	*Drosocarpium*	13	3	3
H. thymifolium	*Taeniocarpium*	18	n/a	2
H. confertum	*Taeniocarpium*	18	n/a	3
H. marginatum	*Taeniocarpium*	18	3	2
H. coris	*Coridium*	19	n/a	2
H. athoum	*Adenosepalum*	27	1	1
H. montanum	*Adenosepalum*	27	3	1
H. delphicum	*Adenosepalum*	27	3	1

6 Discussion

6.1 Comparison of DNA-based Medicinal Plant Identification Techniques

The overall aim of the research introduced in this thesis was to design quick, simple and easy DNA based identification methods which are accessible and advantageous to the medicinal plant industry. The myriad DNA based identification techniques available require the user to clearly assess their requirements before design and/or use of an assay.

Table 6.1 gives a comparison of important factors to be considered when selecting a DNA-based identification technique. The cost of developing a method depends on the time and expertise and, to a large extent, the knowledge required prior to design. Techniques which require DNA sequencing, such as SCAR and ARMS, are more expensive to develop due to this requirement, and barcode sequencing is relatively expensive due to the consumables and technical ability required. The three assays designed in this research have relatively low development costs because they anticipate the availability of freely accessible DNA barcode data as a platform for design.

Running costs are often the factor of most interest to industry and are predominantly determined by the number of stages in the final assay and the consumables used. As PCR-RFLP requires two separate stages it is more time consuming and expensive; the enzyme used can also add substantially to the cost if an unusual one is required. DNA sequencing remains expensive despite advances, and the fluorescently labelled primers and capillary electrophoresis involved in the multiplex assay increase costs, although as only one reaction is needed this is a limited increase.

All of the DNA based methods considered in Table 6.1 are capable of species level identification, as this is the level required by the European Pharmacopoeia for medicinal plant material. Methods with more resolving power can be used for different situations, such as identifying samples according to geographical origin. This can be of particular importance when a border has been crossed and the species in question is protected (see section 6.5).

The final two factors listed in Table 6.1 both refer to the presence of DNA which is not from the target species, and whether the assays are capable of first detecting this DNA and then identifying the species it is from. Techniques which produce a final assay that gives a simple positive or negative result for the presence of the target DNA are not capable of detecting

anything other than this. Those which produce a banding pattern of results may give an indication of the presence of non-target DNAs but this may be difficult to assess due to the random nature of the methods. The most powerful technique for the detection of adulteration or contamination is the fluorescent multiplex PCR method described in Chapter 4, as this is the only method capable of reliably detecting and identifying contaminant DNA.

The fluorescent multiplex PCR assay has many advantageous features, but does have slightly higher running costs than other methods due to the use of fluorescently labelled primers and CE. This method would therefore be recommended for use in particular problem cases, for instance plant species which are known to be mistaken for dangerous alternatives or particular preparations which must contain many different species in order to be efficacious.

The microcode primer design assay is the most cost effective DNA based identification method, with extremely low running and development costs and quick, simple and fast results. With these features this method would be recommended for routine use in batch testing within the supply chain and Quality Assurance monitoring of medicinal plant manufacturers. This could well represent a cost saving to those companies, both in the buying of quality supplies and in the early detection of contaminated, adulterated or misidentified plant material before it has entered into the processing system where it would be much more difficult and expensive to remove.

Table 6.1 Comparison of fundamental factors for different DNA-based identification methods for medicinal plant material, adapted from (Yip et al., 2007).

	RAPD	SCAR	PCR-RFLP	AFLP	ARMS	SSR	ISSR	Barcode Sequencing	Microcode primer design	Microcode qPCR	Multiplex fluorescent PCR
Development cost	low	medium	medium	medium	high	high	low	medium	low	low	low
Running cost	low	low	medium	low	low	low	low	high	low	low	medium
Reliability	low	high	high	high	high	high	medium	high	high	high	high
Species level ID?	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Beyond species level ID?	yes	no	yes	yes	yes	yes	yes	no*	no	no	no
Detection of adulteration/contamination?	yes[1]	no	yes[1]	no	no	yes[1]	yes[1]	yes/no*	no	yes*	yes
Identification of adulterant/contaminant?	n/a	n/a	no	n/a	n/a	n/a	n/a	yes/no*	n/a	no	yes

no*: barcode sequencing can give much more information than species level ID, although it is not designed for this purpose and would not be expected to be used for any reason other than species ID by the medicinal plant industry.

yes/no*: barcode sequencing could detect and identify contaminants and adulterants if the sequence could be separated from the target, if not then neither the target nor the other species could be reliably identified.

yes*: microcode qPCR is capable of detecting contamination or adulteration indirectly, via the different measurements which would result from universal and species-specific qPCR.

yes[1]: These methods could be capable of detecting contaminant or adulterant DNA as a different banding pattern may be produced. However, this would be merged with the target banding pattern so would be difficult to distinguish and separate, meaning these methods are not ideal detecting contaminant or adulterant DNA.

6.2 DNA-based Identification and Hybrids

Plastid regions have been selected for DNA barcoding due to their high copy number, which confers ease of DNA extraction of adequate quality from degraded and processed samples. However, a major criticism of DNA barcoding is the exclusive use of organellar genome regions (Rubinoff et al., 2006) as these are uniparentally inherited.

One important ramification of this is the problem of identification of plant hybrids. Using DNA barcoding, hybrid plants may frequently be incorrectly identified as their maternal parent species as this is usually the plastid donor parent (See section 5.3.5.5). This has been accepted for the present by the DNA barcoding community in order to continue with the barcoding initiative, as they consider the advantages of barcoding identification to greatly outweigh the potential misidentification of hybrids (Chase and Fay, 2009). However, the requirements of the herbal medicine industry may be rather different from those of theoretical plant taxonomy.

The current legislation is aimed at medicinal plants from a designated species, hybrids are not mentioned specifically and therefore the legislation does not directly demand their detection. However, this leads to the problem of distinguishing hybrids from target species when the plastid donor parent of the hybrid is the target species. A possible solution to this is the use of multiple loci from different genomes, including one or many from the biparentally inherited nuclear genome. The nrITS has been shown to be the most useful marker in *Hypericum* species for identification, and is also suited to the detection of hybrids in conjunction with plastid regions. The progression of the design of the multiplex fluorescent PCR method should incorporate multiple loci from nuclear and plastid genomes, all of which produce different peaks which must all be present for a positive ID. This would allow the detection of hybrids by the production of different peak patterns, similar to the banding patterns assessed with AFLP type techniques.

The detection of hybrids is complicated enormously when a sample cannot be guaranteed to be from one plant only, as is the case with the highly processed plant material in herbal medicinal products. The banding or peak patterns produced which designate a hybrid would be almost indecipherable from the patterns produced from a mixed DNA extraction from the two parent species. There may be peak intensity differences which could indicate that each loci measured was not present within the DNA extraction at the expected ratios, which would then require further investigation. In order to successfully and reliably identify hybrids within

the medicinal plant industry by DNA-based assays, the testing for them must be carried out upstream of the processing of the plant material.

6.3 DNA testing of Herbal Medicinal End Products

The types of herbal medicinal end products available to consumers are many and varied. For example, a survey carried out by an MSc student working within this research group found that SJW can be purchased in the form of a pure plant juice, capsules filled with dried plant material, compressed plant material tablets, ethanol extracted tinctures, essential oils and oil containing extracts (Murphy, 2009). The applicability of DNA-based identification methods to these different materials has not been fully investigated. Although it is suggested that many of the DNA-based methods should be employed in the upstream processing stages of manufacture, a test for products as sold over the counter would be a powerful tool for regulators, and therefore consumers, to guarantee the safety and quality of the products on sale.

It has previously been shown that DNA extracted from the plant material used to fill capsules to be sold as herbal medicinal products is of sufficient quality and yield to support DNA-based identification assays (Howard et al., 2009). Tablets of highly compressed plant material with a hard, sugar-type coating also yielded DNA suitable for the microcode PCR assay and gave a positive result for *H. perforatum* when tested as part of the MSc project based on this work (Murphy, 2009). Extracting DNA from tinctures and extracts could prove to be more difficult due to the manufacturing process. However, this has been shown to be possible for a ethanol extract of *Echinacea* sp., and for two different chamomile products: a tincture and a fluid extract (Novak et al., 2007). The DNA extraction was achieved by repeated dilution, centrifugation, and resuspension stages to pellet plant material which remained in the products. This pellet was further washed to remove plant secondary metabolites, and then the DNA was extracted. The final concentration of DNA was too low to be measured, but the nrITS region was amplified in all samples, showing the utility of low copy nuclear regions even in samples where DNA extraction is problematic.

Further work on the SJW microcode assay should continue with the thorough survey of available products and the ability to extract DNA from them that is of suitable quality and quantity to enable DNA-based identification of products. As has been shown, this is possible for several types of end product, even those which yield a very low concentration of DNA. This

would give a guide as to the applicability of DNA-based techniques to end products, and to the quality of those products currently on sale.

6.4 Should DNA-based identification replace chemical methods?

DNA-based techniques are an extremely powerful tool for the fast, effective and unambiguous identification of medicinal plant material and the products derived from it. This is the first fundamental need for quality control and safety within the medicinal plant industry; the starting material must be correct and be shown to be correct at the commencement of the manufacturing process. The major advantages of DNA-based techniques are speed and accuracy, low cost, applicability to all stages of manufacturing and plant life cycle, and flexibility of testing material.

However, DNA-based methods alone cannot be sufficient to maintain quality and safety standards of medicinal plant materials and products, as they cannot measure the quality of plant material with respect to secondary metabolite concentration (Sucher and Carles, 2008). As has been described earlier (section 1.2.2.2) the chemical constituents of plant material are affected by many factors including species, but also harvesting and storage, climate and environment. The analytical methods used to measure these compounds are highly accurate and efficient, and are by far the best means of assessing product quality.

There are several applications in which DNA-based testing may be used as a complement to chemical methods:

- A 'first hurdle', cheap and easy test for companies to ensure that suppliers are delivering the correct plant species.
- Batch testing required for GMP, due to the speed and cost being much lower than chemical methods.
- Regulators requiring spot testing of a large number of samples, enabling focus to fall on the lowest quality products.
- Identification of consumer products which have been "spiked" with the marker chemical compounds tested for, but which do not contain the stated plant species.

The necessity to ensure that the compounds within medicinal plant material are within expected levels precludes the use of DNA-based methods alone. The two types of technique

are complementary and should be used as such to the benefit of the end-user, whether supplier, manufacturer, regulator or retailer.

6.5 Further Applications of Botanical DNA-based Identification Techniques

The initiative of DNA barcoding has attracted much attention and funding worldwide and, due to this, has been required to describe further uses for the information collected and catalogued. The research described in this thesis shows several useful methods of employing these data for the direct use of industry and regulators within the medicinal plant industry.

The medicinal plant arena was chosen for many reasons as detailed in Section 1. However, there are many more arenas which would greatly benefit from a replication of the DNA based assay models developed.

Botanical trace evidence is often found at scenes of crime but has previously not been used to full effect by investigators due to a lack of morphological expertise. This material is also often damaged and degraded, which further complicates morphology based identification. The use of molecular methods to identify plant evidence has increased the accessibility of this evidence to forensic investigators, and the advent of DNA barcoding has increased interest further (Ferri, 2009).

Methods investigated so far have concentrated on regions which were once barcode candidates (*trnH-psbA* and *trnL-trnF*) for 63 plants from the local flora of Moderna, Italy (Ferri, 2009) and on the nrITS and *trnL-trnF* regions for 50 local plants from the Netherlands, with an emphasis on grasses (Wesselink and Kuiper, 2008).

These methods are both based on sequencing of the chosen regions followed by comparison to the GenBank database for species identification, and neither takes advantage of the DNA barcoding data available. The multiplex fluorescent microcode technique could be adapted to several different forensic scenarios. For instance, a method to detect different grass species could be designed, which could then match plant DNA on clothing to the species selection found at a crime scene. The field of Forensics is fast moving, but currently awaits a suitable and reliable method for utilising botanical trace evidence to its full potential (Ferri, 2009).

Within forensics, another area which would benefit from specialised design of molecular identification methods is the trade in endangered species. Enforcement of the Convention on International Trade of Endangered Species (CITES) will be much strengthened by the use of DNA-based techniques to accurately and quickly identify samples.

Molecular techniques have been designed specifically for this purpose; one in order to identify the tropical hardwood ramin, *Gonystylus* spp. This research was conducted in close communication with regulatory bodies in order to ensure the utility of the technique once designed, and resulted in a successful identification technique for both timber and worked material, providing a proof of concept for DNA-based techniques in this arena (Ogden et al., 2008). Further studies similar to this, directed by regulatory bodies to the problem areas and solutions required for them, working together with molecular and other scientists must be the most practical and efficient way to utilise these techniques to their full potential to the benefit of all concerned.

7 References

ABOELSOUD, N. (2010) Herbal Medicine in ancient Egypt. *Journal of Medicinal Plants Research,* 4**,** 082-086.

ALVAREZ, I. & WENDEL, J. F. (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution,* 29**,** 417-434.

AMBION http://www.ambion.com/techlib/tn/102/17.html Accessed 03/04/2010. Applied Biosystems.

APPLIED BIOSYSTEMS, A. (2000) ABI PRISM® SNaPshot™ Multiplex Kit Protocol. *Available at www3.appliedbiosystems.com*.

APTE, A. & SINGH, S. (2007) AlleleID. *PCR Primer Design.*

ASAHINA, H., SHINOZAKI, J., MASUDA, K., MORIMITSU, Y. & SATAKE, M. (2010) Identification of medicinal Dendrobium species by phylogenetic analyses using matK and rbcL sequences. *Journal of Natural Medicines,* 64**,** 133-138.

BANFER, G., MOOG, U., FIALA, B., MOHAMED, M., WEISING, K., BLATTNER, F.R. (2006) A chlororplast genealogy of myrmecophytic Macaranga species (Euphorbiaceae) in Southeat Asia reveals hybridization, vicariance and long-distance dispersals. *Molecular Ecology* 15**,** 4409-24.

BARNES, J. (2003) Quality, efficacy and safety of complementary medicines: fashions, facts and the future. Part I. Regulation and quality. *British Journal of Clinical Pharmacology,* 55**,** 226-233.

BERTEA, C. M., LUCIANO, P., BOSSI, S., LEONI, F., BAIOCCHI, C., MEDANA, C., AZZOLIN, C. M. M., TEMPORALE, G., LOMBARDOZZI, M. A. & MAFFEI, M. E. (2006) PCR and PCR-RFLP of the 5S-rRNA-NTS region and salvinorin A analyses for the rapid and unequivocal determination of Salvia divinorum. *Phytochemistry,* 67**,** 371-378.

BRUTOVSKÁ, R., ČELLÁROVÁ, E. & SCHUBERT, I. (2000) Cytogenetic characterization of three Hypericum species by in situ hybridization. *TAG Theoretical and Applied Genetics,* 101**,** 46-50.

BUSTIN, S. A., BENES, V., GARSON, J. A., HELLEMANS, J., HUGGETT, J., KUBISTA, M., MUELLER, R., NOLAN, T., PFAFFL, M. W., SHIPLEY, G. L., VANDESOMPELE, J. & WITTWER, C. T. (2009) The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry,* 55**,** 611-622.

BUTTERWECK, V. (2003) Mechanism of Action of St John's Wort in Depression: What is Known? *CNS Drugs,* 17**,** 539-562.

BUTTERWECK, V., CHRISTOFFEL, V., NAHRSTEDT, A., PETEREIT, F., SPENGLER, B. & WINTERHOFF, H. (2003) Step by step removal of hyperforin and hypericin: activity profile of different Hypericum preparations in behavioral models. *Life Sciences,* 73**,** 627-639.

BUTTERWECK, V. & DERENDORF, H. (2008) Potential of pharmacokinetic profiling for detecting herbal interactions with drugs. *Clinical Pharmacokinetics,* 47**,** 383-397.

BUTTERWECK, V., JURGENLIEMK, G., NAHRSTEDT, A. & WINTERHOFF, H. (2000) Flavonoids from Hypericum perforatum show antidepressant activity in the forced swimming test. *Planta Medica,* 66**,** 3-6.

BUTTERWECK, V., PETEREIT, F., WINTERHOFF, H. & NAHRSTEDT, A. (1998) Solubilized hypericin and pseudohypericin from Hypericum perforatum exert antidepressant activity in the forced swimming test. *Planta Medica,* 64**,** 291-294.

CARINE, M. A., CHRISTENHUSZ M.J.M. (2010) Editorial. About this volume: the monograph of *Hypericum* by Norman Robson. *Phytotaxa,* 4**,** 1-4.

CBOL, P. W. G., HOLLINGSWORTH, P. M., FORREST, L. L., SPOUGE, J. L., HAJIBABAEI, M., RATNASINGHAM, S., VAN DER BANK, M., CHASE, M. W., COWAN, R. S., ERICKSON, D.

L., FAZEKAS, A. J., GRAHAM, S. W., JAMES, K. E., KIM, K. J., KRESS, W. J., SCHNEIDER, H., VAN ALPHENSTAHL, J., BARRETT, S. C. H., VAN DEN BERG, C., BOGARIN, D., BURGESS, K. S., CAMERON, K. M., CARINE, M., CHACON, J., CLARK, A., CLARKSON, J. J., CONRAD, F., DEVEY, D. S., FORD, C. S., HEDDERSON, T. A. J., HOLLINGSWORTH, M. L., HUSBAND, B. C., KELLY, L. J., KESANAKURTI, P. R., KIM, J. S., KIM, Y. D., LAHAYE, R., LEE, H. L., LONG, D. G., MADRINAN, S., MAURIN, O., MEUSNIER, I., NEWMASTER, S. G., PARK, C. W., PERCY, D. M., PETERSEN, G., RICHARDSON, J. E., SALAZAR, G. A., SAVOLAINEN, V., SEBERG, O., WILKINSON, M. J., YI, D. K. & LITTLE, D. P. (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 12794-12797.

CHASE, M. W., COWAN, R. S., HOLLINGSWORTH, P. M., VAN DEN BERG, C., MADRINAN, S., PETERSEN, G., SEBERG, O., JORGSENSEN, T., CAMERON, K. M., CARINE, M., PEDERSEN, N., HEDDERSON, T. A. J., CONRAD, F., SALAZAR, G. A., RICHARDSON, J. E., HOLLINGSWORTH, M. L., BARRACLOUGH, T. G., KELLY, L. & WILKINSON, M. (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon,* 56**,** 295-299.

CHASE, M. W. & FAY, M. F. (2009) Barcoding of plants and fungi. *Science,* 325**,** 682-683.

CHASE, M. W., SALAMIN, N., WILKINSON, M., DUNWELL, J. M., KESANAKURTHI, R. P., HAIDAR, N. & SAVOLAINEN, V. (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360**,** 1889-1895.

CHATTERJEE, S. S., NOLDNER, M., KOCH, E. & ERDELMEIER, C. (1998) Antidepressant activity of Hypericum perforatum and hyperforin: The neglected possibility. *Pharmacopsychiatry,* 31**,** 7-15.

CHEN, S. L., YAO, H., HAN, J. P., LIU, C., SONG, J. Y., SHI, L. C., ZHU, Y. J., MA, X. Y., GAO, T., PANG, X. H., LUO, K., LI, Y., LI, X. W., JIA, X. C., LIN, Y. L. & LEON, C. (2010) Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. *Plos One,* 5**,** 8.

CHENNA, R., SUGAWARA, H., KOIKE, T., LOPEZ, R., GIBSON, T. J., HIGGINS, D. G. & THOMPSON, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research,* 31**,** 3497-3500.

CHOO, B. K., MOON, B. C., JI, Y., KIM, B. B., CHOI, G., YOON, T. & KIM, H. K. (2009) Development of SCAR Markers for the Discrimination of Three Species of Medicinal Plants, Angelica decursiva (Peucedanum decursivum), Peucedanum praeruptorum and Anthricus sylvestris, Based on the Internal Transcribed Spacer (ITS) Sequence and Random Amplified Polymorphic DNA (RAPD). *Biological & Pharmaceutical Bulletin,* 32**,** 24-30.

CLAUSON, K. A., SANTAMARINA, M. L. & RUTLEDGE, J. C. (2008) Clinically relevant safety issues associated with St. John's wort product labels. *Bmc Complementary and Alternative Medicine,* 8.

CONDON, A. (2006) Designed DNA molecules: principles and applications of molecular nanotechnology. *Nat Rev Genet,* 7**,** 565-575.

COWAN, R. S., CHASE, M. W., KRESS, W. J. & SAVOLAINEN, V. (2006) 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon,* 55**,** 611-616.

CROCKETT, S. L., DEMIRCI, B., BASER, K. H. C. & KHAN, I. A. (2007) Analysis of the volatile constituents of five African and Mediterranean Hypericum L. (Clusiaceae, Hypericoideae) species. *Journal of Essential Oil Research,* 19**,** 302-306.

CROCKETT, S. L., DOUGLAS, A. W., SCHEFFLER, B. E. & KHAN, I. A. (2004) Genetic profiling of Hypericum (St. John's Wort) species by nuclear ribosomal ITS sequence analysis. *Planta Medica,* 70**,** 929-935.

CUENOUD, P., SAVOLAINEN, V., CHATROU, L. W., POWELL, M., GRAYER, R. J. & CHASE, M. W. (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. *American Journal of Botany,* 89**,** 132-144.

DE SMET, P. (2005) Herbal medicine in Europe - Relaxing regulatory standards. *New England Journal of Medicine,* 352**,** 1176-1178.

DEVAIAH, K. M. & VENKATASUBRAMANIAN, P. (2008) Development of SCAR marker for authentication of Pueraria tuberosa (Roxb. ex. Willd.) DC. *Current Science,* 94**,** 1306-1309.

DING, G., ZHANG, D. Z., FENG, Z. Y., FAN, W. J., DING, X. Y. & LI, X. X. (2008) SNP, ARMS and SSH authentication of medicinal Dendrobium officinale KIMURA et MIGo and application for identification of Fengdou drugs. *Biological & Pharmaceutical Bulletin,* 31**,** 553-557.

EBACH, M. C. & HOLDREGE, C. (2005) DNA barcoding is no substitute for taxonomy. *Nature,* 434**,** 697-697.

EUROPEAN DIRECTORATE FOR THE QUALITY OF MEDICINES AND HEALTHCARE (2010) European Pharmacopoeia. *http://www.edqm.eu/en/History-93.html Accessed 05/05/2010*.

EUROPEAN PHARMACOPOEIA (2008) St. John's Wort *Hyperici herba*. *Monograph,* 6.2**,** 3839-3840.

EXECUTIVE COMMITTEE, C. B. O. L. (2009) CBOL approves *matK* and *rbcL* as the BARCODE regions for Land Plants. Consortium for the Barcode of Life.

FAN, L. L., ZHU, S., CHEN, H. B., YANG, D. H., CAI, S. Q. & KOMATSU, K. (2009) Identification of the Botanical Source of Stemonae Radix Based on Polymerase Chain Reaction with Specific Primers and Polymerase Chain Reaction-Restriction Fragment Length Polymorphism. *Biological & Pharmaceutical Bulletin,* 32**,** 1624-1627.

FAZEKAS, A. J., BURGESS, K. S., KESANAKURTI, P. R., GRAHAM, S. W., NEWMASTER, S. G., HUSBAND, B. C., PERCY, D. M., HAJIBABAEI, M. & BARRETT, S. C. H. (2008) Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. *Plos One,* 3.

FELINER, G. N. & ROSSELLO, J. A. (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution,* 44**,** 911-919.

FERRI, G. A., M. CORRADINI, B. AND BEDUSCHI, G. (2009) Forensic Botany: species identification of botanical trace evidence using a multigene barcoding approach. *International Journal of Legal Medicine,* 123**,** 395-401.

FOOD STANDARDS AGENCY, F. S. A. (2005) Guidance Notes on Legislation Implementing Directive 2002/46/EC on Food Supplements. *www.food.gov.uk/multimedia/pdfs/suppsguidancefinal15april05.pdf Accessed 5 May 2010*.

GASTER, B. & HOLROYD, J. (2000) St John's wort for depression - A systematic review. *Archives of Internal Medicine,* 160**,** 152-156.

GREGORY, T. R. (2005) DNA barcoding does not compete with taxonomy. *Nature,* 434**,** 1067-1067.

HANSEN, A. K., ESCOBAR, L. K., GILBERT, L. E. & JANSEN, R. K. (2007) Paternal, maternal, and biparental inheritance of the chloroplast genome in Passiflora (Passifloraceae): Implications for phylogenetic studies. *American Journal of Botany,* 94**,** 42-46.

HEBERT, P. D. N. & GREGORY, T. R. (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology,* 54**,** 852-859.

HEBERT, P. D. N., RATNASINGHAM, S. & DEWAARD, J. R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B-Biological Sciences,* 270**,** S96-S99.

HIGUCHI, R., FOCKLER, C., DOLLINGER, G. & WATSON, R. (1993) KINETIC PCR ANALYSIS - REAL-TIME MONITORING OF DNA AMPLIFICATION REACTIONS. *Bio-Technology,* 11**,** 1026-1030.

HOLLINGSWORTH, P. M., FORREST, L. L., SPOUGE, J. L., HAJIBABAEI, M., RATNASINGHAM, S., VAN DER BANK, M., CHASE, M. W., COWAN, R. S., ERICKSON, D. L., FAZEKAS, A. J., GRAHAM, S. W., JAMES, K. E., KIM, K. J., KRESS, W. J., SCHNEIDER, H., VAN ALPHENSTAHL, J., BARRETT, S. C. H., VAN DEN BERG, C., BOGARIN, D., BURGESS, K. S., CAMERON, K. M., CARINE, M., CHACON, J., CLARK, A., CLARKSON, J. J., CONRAD, F., DEVEY, D. S., FORD, C. S., HEDDERSON, T. A. J., HOLLINGSWORTH, M. L., HUSBAND, B. C., KELLY, L. J., KESANAKURTI, P. R., KIM, J. S., KIM, Y. D., LAHAYE, R., LEE, H. L., LONG, D. G., MADRINAN, S., MAURIN, O., MEUSNIER, I., NEWMASTER, S. G., PARK, C. W., PERCY, D. M., PETERSEN, G., RICHARDSON, J. E., SALAZAR, G. A., SAVOLAINEN, V., SEBERG, O., WILKINSON, M. J., YI, D. K. & LITTLE, D. P. (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 12794-12797.

HOWARD, C., BREMNER, P. D., FOWLER, M. R., ISODO, B., SCOTT, N. W. & SLATER, A. (2009) Molecular Identification of Hypericum perforatum by PCR Amplification of the ITS and 5.8S rDNA Region. *Planta Medica,* 75**,** 864-869.

HUNT, E. J., LESTER, C. E., LESTER, E. A. & TACKETT, R. L. (2001) Effect of St. John's wort on free radical production. *Life Sciences,* 69**,** 181-190.

IRSHAD, S., SINGH, J., KAKKAR, P. & MEHROTRA, S. (2009) Molecular characterization of Desmodium species - An important ingredient of 'Dashmoola' by RAPD analysis. *Fitoterapia,* 80**,** 115-118.

JORDAN, S. A., CUNNINGHAM, D. G. & MARLES, R. J. (2010) Assessment of herbal medicinal products: Challenges, and opportunities to increase the knowledge base for safety assessment. *Toxicology and Applied Pharmacology,* 243**,** 198-216.

JOSHI, K., CHAVAN, P., WARUDE, D. & PATWARDHAN, B. (2004) Molecular markers in herbal drug technology. *Current Science,* 87**,** 159-165.

KASPER, S., VOLZ, H. P., MÖLLER, H. J., DIENEL, A. & KIESER, M. (2008) Continuation and long-term maintenance treatment with Hypericum extract WS® 5570 after recovery from an acute episode of moderate depression -- A double-blind, randomized, placebo controlled long-term trial. *European Neuropsychopharmacology,* 18**,** 803-813.

KHAN, I. A. (2006) Issues related to botanicals. *Life Sciences,* 78**,** 2033-2038.

KOBER, M., POHL, K. & EFFERTH, T. (2008) Molecular Mechanisms Underlying St. John's Wort Drug Interactions. *Current Drug Metabolism,* 9**,** 1027-1037.

KOCH, W. H. (2004) Technology platforms for pharmacogenomic diagnostic assays. *Nat Rev Drug Discov,* 3**,** 749-761.

KRESS, W. J. & ERICKSON, D. L. (2007) A Two-Locus Global DNA Barcode for Land Plants: The Coding <italic>rbcL</italic> Gene Complements the Non-Coding <italic>trnH-psbA</italic> Spacer Region. *Plos One,* 2**,** e508.

KRESS, W. J., WURDACK, K. J., ZIMMER, E. A., WEIGT, L. A. & JANZEN, D. H. (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 8369-8374.

KUBISTA, M. (2008) Emerginf real-time PCR applications. *Drug Discovery World,* 9**,** 57-66.

LAHAYE, R., VAN DER BANK, M., BOGARIN, D., WARNER, J., PUPULIN, F., GIGOT, G., MAURIN, O., DUTHOIT, S., BARRACLOUGH, T. G. & SAVOLAINEN, V. (2008) DNA barcoding the

floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 2923-2928.
LI, X. X., DING, X. Y., CHU, B. H., DING, G., GU, S., QIAN, L., WANG, Y. & ZHOU, Q. (2007) Molecular authentication of Alisma orientale by PCR-RFLP and ARMS. *Planta Medica,* 73**,** 67-70.
LINDE, K., BERNER, MM., KRISTON, L. (2008) St John's Wort for major depression. *Cochrane Database of Systematic Reviews*.
LIVAK, K. J., FLOOD, S. J. A., MARMARO, J., GIUSTI, W. & DEETZ, K. (1995) OLIGONUCLEOTIDES WITH FLUORESCENT DYES AT OPPOSITE ENDS PROVIDE A QUENCHED PROBE SYSTEM USEFUL FOR DETECTING PCR PRODUCT AND NUCLEIC-ACID HYBRIDIZATION. *Pcr-Methods and Applications,* 4**,** 357-362.
LIVAK, K. J. & SCHMITTGEN, T. D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. *Methods,* 25**,** 402-408.
LOU, M. & GOLDING, G. B. (2007) FINGERPRINT: visual depiction of variation in multiple sequence alignments. *Molecular Ecology Notes,* 7**,** 908-914.
MARSTON, A. & HOSTETTMANN, K. (2009) Natural Product Analysis over the Last Decades. *Planta Med,* 75**,** 672-682.
MCGARRY, H., PIROTTA, M., HEGARTY, K. & GUNN, J. (2007) General practitioners and St. John's wort: A question of regulation or knowledge? *Complementary Therapies in Medicine,* 15**,** 142-148.
MEDICINES AND HEALTHCARE PRODUCTS REGULATORY AGENCY, M. H. R. A. (2009) Ipsos MORI report results on herbal medicines. *Press Release*.
MEUSNIER, I., SINGER, G., LANDRY, J.-F., HICKEY, D., HEBERT, P. & HAJIBABAEI, M. (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics,* 9**,** 214.
MOON, B. C., CHOO, B. K., CHEON, M. S., YOON, T., JI, Y., KIM, B. B., LEE, A. Y. & KIM, H. K. (2010) Rapid molecular authentication of three medicinal plant species, Cynanchum wilfordii, Cynanchum auriculatum, and Polygonum multiflorum (Fallopia multiflorum), by the development of RAPD-derived SCAR markers and multiplex-PCR. *Plant Biotechnology Reports,* 4**,** 1-7.
MOORE, L. B., GOODWIN, B., JONES, S. A., WISELY, G. B., SERABJIT-SINGH, C. J., WILLSON, T. M., COLLINS, J. L. & KLIEWER, S. A. (2000) St. John's wort induces hepatic drug metabolism through activation of the pregnane X receptor. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 7500-7502.
MURPHY, O. (2009) DNA-based authentication of Hypericum perforatum L. based medicines. *Faculty of Science.* Staffordshire, University of Staffordshire.
NATIONAL INSTITUTES OF HEALTH, N. C. C. A. M. (2008) The use of Complementary and Alternative Medicine in the United States. *CDC National Health Statistics Report* 12.
NEWMASTER, S. G., FAZEKAS, A. J., STEEVES, R. A. D. & JANOVEC, J. (2008) Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources,* 8**,** 480-490.
NIELSEN, K., MOGENSEN, H. S., HEDMAN, J., NIEDERSTATTER, H., PARSON, W. & MORLING, N. (2008) Comparison of five DNA quantification methods. *Forensic Science International-Genetics,* 2**,** 226-230.
NOVAK, J., GRAUSGRUBER-GROGER, S. & LUKAS, B. (2007) DNA-based authentication of plant extracts. *Food Research International,* 40**,** 388-392.
OGDEN, R., MCGOUGH, H. N., COWAN, R. S., CHUA, L., GROVES, M. & MCEWING, R. (2008) SNP-based method for the genetic identification of ramin Gonystylus spp. timber and products: applied research meeting CITES enforcement needs. *Endangered Species Research,* 9**,** 255-261.

PARK, S.-J. & KIM, K.-J. (2004) Molecular phylogeny of the genus Hypericum (Hypericaceae) from Korea and Japan: evidence from nuclear rDNA ITS sequence data. *Journal of Plant Biology,* 47**,** 366-374.

PERCIFIELD, R. J., HAWKINS, J. S., MCCOYZ, J. A., WIDRLECHNERZ, M. P. & WENDEL, J. F. (2007) Genetic diversity in Hypericum and AFLP markers for species-specific identification of H. perforatum L. *Planta Medica,* 73**,** 1614-1621.

PFAFFL, M. W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research,* 29.

PREMIERBIOSOFT (2010) http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html, accessed 19/04/2010.

PROMEGA (2009) http://www.promega.com/paguide/chap1.htm, accessed 31/03/2010.

QIAGEN (2006) *DNeasy Plant Handbook*, Qiagen Group.

RADUSIENE, J., JUDZENTIENE, A. & BERNOTIENE, G. (2005) Essential oil composition and variability of Hypericum perforatum L. growing in Lithuania. *Biochemical Systematics and Ecology,* 33**,** 113-124.

RAHIMI, R., NIKFAR, S. & ABDOLLAHI, M. (2009) Efficacy and tolerability of Hypericum perforatum in major depressive disorder in comparison with selective serotonin reuptake inhibitors: A meta-analysis. *Progress in Neuro-Psychopharmacology and Biological Psychiatry,* 33**,** 118-127.

REBRIKOV, D. V. & TROFIMOV, D. Y. (2006) Real-time PCR: A review of approaches to data analysis. *Applied Biochemistry and Microbiology,* 42**,** 455-463.

ROBSON, N. (2002) Studies in the genus *Hypericum* L. (Guttiferae) 4(2). Section 9. *Hypericum* sensu lato (part 2): Subsection 1. *Hypericum* series 1. *Hypericum*. *Bulletins of the Natural History Museum London, Botany,* 32**,** 61-123.

ROBSON, N. K. B. (2006) Studies in the genus Hypericum L. (Clusiaceae). Section 9. Hypericum sensu lato (part 3): subsection 1. Hypericum series 2. Senanensia, subsection 2. Erecta and section 9b. Graveolentia. *Systematics and Biodiversity,* 4**,** 19-98.

ROZEN S., S. H. J. (2000) *Primer3 on the WWW for general users and for biologist programmers,* Totowa NJ, Humana Press.

RUBINOFF, D., CAMERON, S. & WILL, K. (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of Heredity,* 97**,** 581-594.

RUSSI, L., MORETTI, C., RAGGI, L., ALBERTINI, E. & FALISTOCCO, E. (2009) Identifying commercially relevant Echinacea species by AFLP molecular markers. *Genome,* 52**,** 912-918.

SASS, C., LITTLE, D. P., STEVENSON, D. W. & SPECHT, C. D. (2007) DNA Barcoding in the Cycadales: Testing the Potential of Proposed Barcoding Markers for Species Identification of Cycads. *Plos One,* 2.

SAXENA, A., PARIJAT TRIPATHI, K., ROY, S., KHAN, F. & SHARMA, A. (2008) Pharmacovigilance: Effects of herbal components on human drugs interactions involving Cytochrome P450. *Bioinformation,* 3**,** 198-204.

SCHINDEL, D. E. & MILLER, S. E. (2005) DNA barcoding a useful tool for taxonomists. *Nature,* 435**,** 17-17.

SMELCEROVIC, A., VERMA, V., SPITELLER, M., AHMAD, S. M., PURI, S. C. & QAZI, G. N. (2006) Phytochemical analysis and genetic characterization of six Hypericum species from Serbia. *Phytochemistry,* 67**,** 171-177.

SMITH, G. F. & FIGUEIREDO, E. (2009) Capacity building in taxonomy and systematics. *Taxon,* 58**,** 697-699.

SOUTHWELL, I. A. & BOURKE, C. A. (2001) Seasonal variation in hypericin content of Hypericum perforatum L. (St. John's Wort). *Phytochemistry,* 56**,** 437-441.

SUCHER, N. J. & CARLES, M. C. (2008) Genome-based approaches to the authentication of medicinal plants. *Planta Medica,* 74**,** 603-623.

SUKRONG, S., ZHU, S., RUANGRUNGSI, N., PHADUNGCHAROEN, T., PALANUVEJ, C. & KOMATSU, K. (2007) Molecular analysis of the genus Mitragyna existing in Thailand based on rDNA ITS sequences and its application to identify a narcotic species: Mitragyna speciosa. *Biological & Pharmaceutical Bulletin,* 30**,** 1284-1288.

SUMMERBELL, R. C., LEVESQUE, C. A., SEIFERT, K. A., BOVERS, M., FELL, J. W., DIAZ, M. R., BOEKHOUT, T., DE HOOG, G. S., STALPERS, J. & CROUS, P. W. (2005) Microcoding: the second step in DNA barcoding. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360**,** 1897-1903.

TECHEN, N., CROCKETT, S. L., KHAN, I. A. & SCHEFFLER, B. E. (2004) Authentication of medicinal plants using molecular biology techniques to compliment conventional methods. *Curr Medicinal Chem,* 11**,** 1391-1401.

TECHEN, N., PAN, Z. Q., SCHEFFIER, B. E. & KHAN, I. A. (2009) Detection of Illicium anisatum as Adulterant of Illicium verum. *Planta Medica,* 75**,** 392-395.

THAKKAR, N. V. & PATEL, J. A. (2010) Pharmacological evaluation of "Glyoherb": A polyherbal formulation on streptozotocin-induced diabetic rats. *International Journal of Diabetes in Developing Countries,* 30**,** 1-7.

TOBE, S. S. & LINACRE, A. M. T. (2008) A multiplex assay to identify 18 European mammal species from mixtures using the mitochondrial cytochrome b gene. *Electrophoresis,* 29**,** 340-347.

TRIPATHI, Y. B., SINGH, B. K., PANDEY, R. S. & KUMAR, M. (2005) BHUx: A Patent Polyherbal Formulation to Prevent Atherosclerosis. *eCAM,* 2**,** 217-221.

TYAGI, S. & KRAMER, F. R. (1996) Molecular Beacons: Probes that Fluoresce upon Hybridization. *Nat Biotech,* 14**,** 303-308.

UJIHARA, T., OHTA, R., HAYASHI, N., KOIATA, K. & TANAKA, J. I. (2009) Identification of Japanese and Chinese Green Tea Cultivars by Using Simple Sequence Repeat Markers to Encourage Proper Labeling. *Bioscience Biotechnology and Biochemistry,* 73**,** 15-20.

UK HOUSE OF COMMONS (2003) Debate. *Hansard,* c191W.

VALLONE, P. M., BUTLER, J.M. (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques,* 37**,** 226-31.

VANHERWEGHEM, J. (1998) Misuse of herbal remedies: the case of an outbreak of terminal renal failure in Belgium (Chinese herbs nephropathy)... including commentary by McIntyre M. *Journal of Alternative & Complementary Medicine,* 4**,** 9-16.

VERMEULEN, J., PATTYN, F., DE PRETER, K., VERCRUYSSE, L., DERVEAUX, S., MESTDAGH, P., LEFEVER, S., HELLEMANS, J., SPELEMAN, F. & VANDESOMPELE, J. (2009) External oligonucleotide standards enable cross laboratory comparison and exchange of real-time quantitative PCR data. *Nucleic Acids Research,* 37**,** 9.

VLIETINCK, A., PIETERS, L. & APERS, S. (2009) Legal Requirements for the Quality of Herbal Substances and Herbal Preparations for the Manufacturing of Herbal Medicinal Products in the European Union. *Planta Med,* 75**,** 683-688.

WAGNER, H. (2009) Synergy Research: a New Approach to Evaluating the Efficacy of Herbal Mono-drug Extracts and Their Combinations. *Natural Product Communications,* 4**,** 303-304.

WALLACE, R. B., SHAFFER, J., MURPHY, R. F., BONNER, J., HIROSE, T. & ITAKURA, K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to {Phi}X 174 DNA: the effect of single base pair mismatch. *Nucl. Acids Res.,* 6**,** 3543-3558.

WANG, H., SUN, H., KWON, W.-S., JIN, H. & YANG, D.-C. (2010) A PCR-based SNP marker for specific authentication of Korean ginseng (panax ginseng) cultivar "Chunpoong". *Molecular Biology Reports,* 37**,** 1053-1057.

WANG, H. Z., FENG, S. G., LU, J. J., SHI, N. N. & LIU, J. J. (2009) Phylogenetic study and molecular identification of 31 Dendrobium species using inter-simple sequence repeat (ISSR) markers. *Scientia Horticulturae,* 122**,** 440-447.

WANG, Q., CHENG, Z., ZHANG, L., WAN, S.-W., DING, J.-M., FU, D.-X., CHEN, J.-K. & ZHANG, W.-J. (2004) Distinguishing wild Panax ginseng from cultivated Panax ginseng based on direct amplification of length polymorphism (DALP) analysis. *Fudan Xuebao Zirankexueban,* 43**,** 1030-1034.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M. A., CLAMP, M. & BARTON, G. J. (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-1191.

WESSELINK, M. & KUIPER, I. (2008) Species identification of botanical trace evidence using molecular markers. *Forensic Science International: Genetics Supplement Series,* 1**,** 630-632.

WHITE, T. J., BRUNS, T., LEE, S. & TAYLOR, J. (1990) *AMPLIFICATION AND DIRECT SEQUENCING OF FUNGAL RIBOSOMAL RNA GENES FOR PHYLOGENETICS*.

WILHELM, J. & PINGOUD, A. (2003) Real-time polymerase chain reaction. *Chembiochem,* 4**,** 1120-1128.

WILL-SHAHAB, L., BAUER, S., KUNTER, U., ROOTS, I. & BRATTSTRÖM, A. (2009) St John's wort extract (Ze 117) does not alter the pharmacokinetics of a low-dose oral contraceptive. *European Journal of Clinical Pharmacology,* 65**,** 287-294.

WILLIAMSON, E. M. (2001) Synergy and other interactions in phytomedicines. *Phytomedicine,* 8**,** 401-409.

WORLD HEALTH ORGANIZATION, W. H. O. (2010) Traditional Medicine. *Fact sheet,* 134.

XU, H., FABRICANT, D. S., PIERSEN, C. E., BOLTON, J. L., PEZZUTO, J. A., FONG, H., TOTURA, S., FARNSWORTH, N. R. & CONSTANTINOU, A. I. (2002) A preliminary RAPD-PCR analysis of Cimicifuga species and other botanicals used for women's health. *Phytomedicine,* 9**,** 757-762.

YIP, P. Y., CHAU, C. F., MAK, C. Y. & KWAN, H. S. (2007) DNA methods for identification of Chinese medicinal materials. *Chin Med,* 2**,** 9.

ZEREGA, N. J. C., MORI, S., LINDQVIST, C., ZHENG, Q. Y. & MOTLEY, T. J. (2002) Using amplified fragment length polymorphisms (AFLP) to identify black cohosh (Actaea Racemosa). *Economic Botany,* 56**,** 154-164.

ZHA, X. Q., LUO, J. P., WANG, J. H., WEI, Z. J. & JIANG, S. T. (2009) Genetic characterization of the nine medicinal Dendrobium species using RAPD. *African Journal of Biotechnology,* 8**,** 2064-2068.

8 Appendices

8.1.1 Hypericum nrITS sequence species and GenBank Accession Numbers

EF015304 *H. canariense var. floribundum;* EF015303 *H. canariense var. canariense;* DQ006013 *H. mutilum;* AY573026 *H. boreale;* AY573025 *H. japonicum;* AY573024 *H. laxum;* AY573023 *H. jeongjocksanense;* AY573022 *H. ternum;* AY573021 *H. rigidum;* AY573020 *H. setosum;* AY573019 *H. brevistylum;* AY573018 *H. formosum;* AY573017 *H. scoreri;* AY573016 *H. vulcanicum;* AY573015 *H. ascyron;* AY573014 *H. ascyron;* AY573013 *H. aegypticum;* AY573012 *H. androsaemum;* AY573011 *H. sampsonii;* AY573010 *H. leschenaultii;* AY573009 *H. delpinicum;* AY573008 *H. olympicum;* AY573007 *H. maculatum;* AY573006 *H. triquetrifolium;* AY573005 *H. oliganthum;* AY573004 *H. yezoense;* AY573003 *H. oaxacum;* AY573002 *H. pseudopetiolatum;* AY573001 *H. kinashianum;* AY573000 *H. hakonense;* AY572999 *H. sikokumontanum;* AY572998 *H. ovalifolium;* AY572997 *H. asahinae;* AY572996 *H. chejuense;* AY572995 *H. attenuatum;* AY572994 *H. vaniotii;* AY572993 *H. attenuatum;* AY572992 *H. kamtschaticum;* AY572991 *H. erectum;* AY555889 *H. dolabriforme;* AY555888 *H. nudiflorum;* AY555887 *H. frondosum;* AY555886 *H. densiflorum; AY555885 H. lissophloeus;* AY555884 *H. cerastoides;* AY555883 *H. apocynifolium;* AY555882 *H. tetrapetalum;* AY555881 *H. cistifolium;* AY555880 *H. buckleyi;* AY555879 *H. hypericoides;* AY555878 *H. sphaerocarpum;* AY555877 *H. microsepalum;* AY555876 *H. lobocarpum;* AY555875 *H. myrtifolium;* AY555874 *H. crux-andreae;* AY555873 *H. prolificum;* AY555872 *H. tenuifolium;* AY555871 *H. nitidum subsp. nitidum;* AY555870 *H. brachyphyllum;* AY555869 *H. chapmanii;* AY555868 *H. fasciculatum;* AY555867 *H. lloydii;* AY555866 *H. sp.* SLC; AY555865 *H. adpressum;* AY555864 *H. galioides;* AY555863 *H. roeperianum;* AY555862 *H. balearicum;* AY555861 *H. calycinum;* AY555860 *H. patulum;* AY555859 *H. henryi subsp. uraloides;* AY555858 *H. forrestii;* AY555857 *H. leschenaultii;* AY555856 *H. choisianum;* AY555855 *H. x moserianum;* AY555854 *H. lancasteri;* AY555853 *H. kouytchense;* AY555852 *H. beanii;* AY555851 *H. acmosepalum;* AY555850 *H. pseudohenryi;* AY555849 *H. ascyron; AY555848 H. pallens;* AY555847 *H. ericoides;* AY555846 *H. athoum;* AY555845 *H. delphicum;* AY555844 *H. punctatum;* AY555843 *H. graveolens;* AY555841 *H. perforatum subsp. veronense;* AY555840 *H. perforatum subsp. perforatum;* AY555839 *H. perforatum subsp. perforatum;* AJ414728 *H. calycinum;* AF455674 *H. perforatum.*

8.1.2 PCR Primers Designed for Multiplex PlantID System

Primer name	Sense Primer Sequence, 5′ to 3′
Hper.F.1.1	TGTAACGCTCCCGGCTGTG
Hatt.F.1.1	TCCCGTGCGCTCCCATTC
Hath.F.1.1	CCCCGAAATTCCGATATCTC
Hasc.F.1.1	GCTTTCCTTCGGTTCATAAC
Hkou.F.1.1	GGTGGCGGTCAGGCGTGCCAAGCTC
Hand.F.1.1	GCGGCTGTCCTCCTGTTC
Hasc.F.2.1	CTTTCCTTCGGTTCATAAC
Hcal.F.2.1	GTGGCTTTCCTTCTGTTC
Hmac.F.2.1	GGGGCTTCCTTCTGTTCATAAC
Hper.F.3.1	AGGCGTGCCAAGCTCTTG
Hatt.F.3.1	GCATCATAAGAAGTGTTTGG
Hper.F.4.1	GGGGCTTCCTTCTGTTCATAAC
Hper.F.1.2	ATAAGAAGTGTAACGCTCCCGGCTGTG
Hper.F.1.3	GAAGTGTAACGCTCCCGGCTGTG
Hper.F.1.4	AGTGTAACGCTCCCGGCTGTG
Hatt.F.1.2	GAACTTTTGCATCATAAGAAGTGTTTGG
Hatt.F.1.3	GCATCATAAGAAGTGTTTGGCTCAC
Hath.F.1.2	TGGGTGTCACACATCGTTGCC
Hath.F.1.3	GTGTCACACATCGTTGCC
Hath.F.1.4	GGTGTCACACATCGTTGCC
Hasc.F.1.2	TTCCTTCGGTTCATAACTAAAAC
Hasc.F.1.4	GTGGCTTTCCTTCGGTTC
Hasc.F.1.3	TTTCCTTCGGTTCATAACTAAAAC
Hasc.F.1.5	TCCTTCGGTTCATAACTAAAAC
Hand.F.1.2	ACATCGTCGCCCCAAACC
Hand.F.1.3	AAATGTGATACTTGGTGTGAATTGC
Hand.F.1.4	CACATCGTCGCCCCAAAC
Hand.F.1.5	CGGCTGTCCTCCTGTTCATAAC
Hasc.F.2.2	GGTTCATAACTAAAACGACTCTC
Hasc.F.2.3	TCATAACTAAAACGACTCTC
Hasc.F.2.4	CATGAGAAGGACAATGCC
Hcal.F.2.2	GGCTTTCCTTCTGTTCATAAC
Hcal.F.2.3	GAGTTTATTGGGTATAGGATGG
Hcal.F.2.4	TCATAACCAAAACGACTCTC
Hand.F.2.1	CGAAATGTGATACTTGGTGTGAATTG
Hand.F.2.2	CACATCGTCGCCCCAAAC
Hmac.F.2.2	GTCGGGGGCTTCCTTCTG
Hmac.F.2.3	CGGGGGCTTCCTTCTGTTC

Primer name	Anti-sense Primer, 5′ to 3′
Hper.R.1.1	CCGATTGTCTCTTGCGAGATATC
Hatt.R.1.1	ATGTCCCTAAGAGCAATGCAAGGC
Hath.R.1.1	CTTACAACCACCGCTAGTC
Hasc.R.1.1	GAACTCGAAAGAGGCATTG
Hkou.R.1.1	AGTCGTTTTAGTTATGAACAGAAGGAAAG
Hand.R.1.1	GCAATTCACACCAAGTATCACATTTCG
Hasc.R.2.1	AAAGAGGCATTGGTTTTG
Hcal.R.2.1	CCATCCTATACCCAATAAACTC
Hmac.R.2.1	ATCCTATTCCCGATTGTGTCTTG
Hper.R.3.1	GCCGGGGGTTTGTTTGTTG
Hatt.R.3.1	ATCCTCTTCCCGATTGTC
Hper.R.4.1	TCTTGCGAGATATCGGGATTTTG
Hatt.R.1.3	TCCCTAAGAGCAATGCAAGGCTCACGAC
Hatt.R.1.2	GGGCCAACCGCGAATGGG
Hatt.R.1.4	GGTCAACATGTCCCTAAGAGCAATG
Hath.R.1.2	CCGCTAGTCGTGGCTTTGCTTTG
Hath.R.1.4	CCGCTAGTCGTGGCTTTG
Hath.R.1.3	CAACCACCGCTAGTCGTG
Hasc.R.1.2	ACCCAATGAACTCGAAAGAG
Hkou.R.1.2	GTTTTAGTTATGAACAGAAGGAAAGCC
Hkou.R.1.3	GAGAGTCGTTTTAGTTATGAACAGAAGG
Hand.R.1.2	CCATTATCCGCCCCATCCTC
Hand.R.1.3	CGAGGTGTTGGGTTTGGG
Hand.R.1.4	ACCATTATCCGCCCCATCC
Hand.R.1.5	TTATCCGCCCCATCCTCTTC
Hand.R.1.6	TCACACCAAGTATCACATTTCGCTAC
Hasc.R.2.2	CCCATCCTGTACCCAATG
Hasc.R.2.3	TATCCGCCCATCCTGTAC
Hasc.R.2.4	ATCCTGTACCCAATGAAC
Hasc.R.2.5	TTATGAACCGAAGGAAAGC
Hcal.R.2.2	ACTATAACCGAGGGTCTTAC
Hcal.R.2.3	CCCATCCTATACCCAATAAAC
Hmac.R.2.2	CCGATTGTGTCTTGCGAGATATC
Hmac.R.2.3	TATTCCCGATTGTGTCTTGCG

8.1.3 Multiplex PlantID System Primer Testing Results

Forward Primer name				Tm	Reverse Primer name				Tm	Product Length	Target product?	Gel	Panel product?	Gel	Target Product @ correct conc. + 64°C	Gel	Candidate @ 64°C?	25th Sep 63°C Results	Gel	Candidate @ 63°C?
Hand	F	1	1	67	Hand	R	1	1	67.4	104	Y - Genomic	2009-08-20 17.36	N @ 67	2009-09-11 17.08				N/A		
Hand	F	1	2	67	Hand	R	1	2	66.5	65	Y - Genomic	2009-08-20 17.36	N @ 64	2009-09-15 15.43/6	Y	2009-09-25 16.44	Y	No Panel Product	2009-09-28 15.41	Y
Hand	F	1	3	64.7	Hand	R	1	3	64.8	135	Y - Genomic	2009-08-20 17.36	N @ 64	2009-09-24 12.23	Y	2009-09-25 16.44	Y	No Panel Product	2009-09-28 15.41	Y
Hand	F	1	4	65.6	Hand	R	1	4	65.9	67	Y - Genomic	2009-08-20 17.36	N @ 64	2009-09-24 12.23	Y	2009-09-25 16.44	Y	No Panel Product	2009-09-28 15.41	Y
Hand	F	1	4	65.6	Hand	R	1	5	65.4	63	Y - Genomic	2009-08-20 17.36	N @ 64	2009-09-15 15.43/6	Y	2009-09-25 16.44	Y	No Panel Product	2009-09-28 15.41	Y
Hand	F	1	5	67.1	Hand	R	1	6	67.3	98	Y - Genomic	2009-08-20 17.36	N @ 64	2009-09-24 12.23	Y	2009-09-25 16.44	Y	No Panel Product	2009-09-28 15.41	Y
Hand	F	2	1	65.1	Hand	R	1	3	64.8	137	Y - Genomic	2009-08-20 17.36	N @ 64	2009-09-24 12.23	Y	2009-09-25 16.44	Y	No Panel Product	2009-09-28 15.41	Y
Hasc	F	1	1	60.9	Hasc	R	1	1	60.4	221	Y - ABI ITS 12.5	2009-09-16 15.22								
Hasc	F	2	1	57.5	Hasc	R	2	1	57.5	213	Y - ABI ITS 12.5	2009-09-16 15.22	N @ 62	2009-09-17 17.19	N	2009-09-25 16.44		No Target Product	2009-09-28 15.41	N
Hasc	F	1	2	60.9	Hasc	R	1	2	61.7	225	Y - ABI ITS 12.5	2009-09-16 15.22	N @ 62	2009-09-17 17.19	N	2009-09-25 16.44	N	Faint Target Product	2009-09-28 15.41	Y
Hasc	F	1	4	62.5	Hasc	R	1	2	61.7	231	Y - ABI ITS 12.5	2009-09-16 15.22	N @ 62	2009-09-17 17.19	Y	2009-09-25 16.44	Y	Target Product	2009-09-28 15.41	Y
Hasc	F	1	4	62.5	Hasc	R	1	1	60.4	224	Y - ABI ITS 12.5	2009-09-16 15.22								
Hasc	F	1	3	61.5	Hasc	R	1	1	60.4	219	Y - ABI ITS 12.5	2009-09-16 15.22	N @ 62	2009-09-17 17.19	N	2009-09-25 16.44	N	Faint Target Product	2009-09-28 15.41	Y
Hasc	F	1	5	60.2	Hasc	R	1	1	60.4	217	Y - ABI ITS 12.5	2009-09-16 15.22	N @ 62	2009-09-17 17.19	N	2009-09-25 16.44	N	Faint Target Product	2009-09-28 15.41	Y
Hasc	F	2	2	61	Hasc	R	2	2	60.9	228	Y - ABI ITS 12.5	2009-09-16 15.22								
Hasc	F	2	2	61	Hasc	R	2	3	62.1	234	Y - ABI ITS 12.5	2009-09-16 15.22								
Hasc	F	2	3	56.2	Hasc	R	2	4	57.7	222	Y - ABI ITS 12.5	2009-09-16 15.22	N @ 62	2009-09-17 17.19	N	2009-09-25 16.44		No Target Product	2009-09-28 15.41	N
Hasc	F	2	4	59.6	Hasc	R	2	5	59.3	73	Y - ABI ITS 12.5	2009-09-16 15.22	N @ 62	2009-09-17 17.19	N	2009-09-25 16.44		No Target Product	2009-09-28 15.41	N
Hasc	F	1	1	60.9	Hasc	R	2	2	60.9	238	Y - ABI ITS 12.5	2009-09-16 15.22								
Hath	F	1	1	61.8	Hath	R	1	1	61.7	137	Y - ABI ITS 8.3	2009-09-10 15.22	N @ 60	2009-09-16 15.22	Y	2009-09-25 16.44	Y			Y- in theory
Hath	F	1	1	61.8	Hath	R	1	3	64.6	133	Y - ABI ITS 8.3	2009-09-10 15.23	N @ 60	2009-09-16 15.23	Y	2009-09-25 16.44	Y			Y- in theory
Hath	F	1	2	69.7	Hath	R	1	2	70.7	148	Y - ABI ITS 8.3	2009-09-10 15.24						Target and Panel Produ	2009-09-28 15.41	N
Hath	F	1	1	61.8	Hath	R	1	4	64.9	127	N		N @ 62	2009-09-17 17.19	Y	2009-09-25 16.44	Y			Y- in theory
Hath	F	1	3	63.2	Hath	R	1	3	64.6	151	Y - ABI ITS 8.3	2009-09-10 15.24					N	arget, No Panel Produ	2009-09-28 15.41	Y
Hath	F	1	4	65.8	Hath	R	1	4	64.9	146	N							Target and Panel Produ	2009-09-28 15.41	N
Hcal	F	2	1	59.7	Hcal	R	2	1	59.6	240	Y - Genomic	2009-08-20 17.36	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.56	N			N - previous results
Hcal	F	2	2	61	Hcal	R	2	1	59.6	238	Y - Genomic	2009-08-20 17.36	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.57	N			N - previous results
Hcal	F	2	3	59.6	Hcal	R	2	2	59.7	128	Y - Genomic	2009-08-20 17.36	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.58	N			N - previous results
Hcal	F	2	4	59.2	Hcal	R	2	1	59.6	224	Y - Genomic	2009-08-20 17.36	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.59	N			N - previous results
Hcal	F	2	4	59.2	Hcal	R	2	3	59.4	225	Y - Genomic	2009-08-20 17.36	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.60	N			N - previous results
Hcal	F	2	1	59.7	Hcal	R	2	3	59.4	241	Y - Genomic	2009-08-20 17.36	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.61	N			N - previous results
Hkou	F	1	1	79.8	Hkou	R	1	1	65.8	179	Y - Genomic	2009-08-20 17.36	Y @ 62	2009-09-18 16.28	Y	2009-09-24 16.56	N			N - previous results
Hkou	F	1	1	79.8	Hkou	R	1	2	64.9	175	Y - Genomic	2009-08-20 17.36	Y @ 62	2009-09-18 16.28	Y	2009-09-24 16.56	N			N - previous results
Hkou	F	1	1	79.8	Hkou	R	1	3	65.7	182	Y - Genomic	2009-08-20 17.36	Y @ 67	2009-09-11 17.08	N/A	N/A	N			N - previous results
Hmac	F	2	1	65.1	Hmac	R	2	1	64.4	240	Y - Genomic	2009-08-24 17.21	N @ 62	2009-09-16 16.42	N	2009-09-25 16.44	N	No Target, No Panel	2009-09-28 15.41	N
Hmac	F	2	1	65.1	Hmac	R	2	2	65.5	231	Y - Genomic	2009-08-24 17.21	N @ 62	2009-09-16 16.42	N	2009-09-25 16.44	N	No Target, No Panel	2009-09-28 15.41	N
Hmac	F	2	1	65.1	Hmac	R	2	3	64.9	236	Y - Genomic	2009-08-24 17.21	N @ 62	2009-09-16 16.42	N	2009-09-25 16.44	N	No Target, No Panel	2009-09-28 15.41	N
Hmac	F	2	2	66.4	Hmac	R	2	2	65.5	235	N	2009-08-24 17.21	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.56	N	No Target, No Panel	2009-09-28 15.41	N
Hmac	F	2	3	66.9	Hmac	R	2	1	64.4	242	Y - Genomic	2009-08-24 17.21	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.56	N	No Target, No Panel	2009-09-28 15.41	N
Hmac	F	2	3	66.9	Hmac	R	2	2	65.5	233	Y - Genomic	2009-08-24 17.21	Y @ 62	2009-09-22 16.40	N	2009-09-24 16.56	N	No Target, No Panel	2009-09-28 15.41	N
Hper	F	1	1	69	Hper	R	1	1	65	273	N	2009-08-19 15.54	N @ 62	2009-09-16 16.42	Y	2009-09-25 16.44	Y			Y- in theory
Hper	F	3	1	67.3	Hper	R	3	1	66.7	60	Pale Band - Genomic	2009-08-19 15.54								
Hper	F	4	1	65.1	Hper	R	4	1	64.9	222	Y - Genomic	2009-08-19 15.54	N @ 62	2009-09-16 16.42	Y	2009-09-25 16.44	Y			Y- in theory
Hper	F	1	2	72.5	Hper	R	1	1	65	281	Y - Genomic	2009-08-19 15.54	N @ 62	2009-09-16 16.42	Y	2009-09-25 16.44	Y			Y- in theory
Hper	F	1	3	71.9	Hper	R	1	1	65	277	Y - Genomic	2009-08-19 15.54	N @ 62	2009-09-22 16.40	Y	2009-09-25 16.44	Y			Y- in theory
Hper	F	1	4	71.2	Hper	R	1	1	65	275	Y - Genomic	2009-08-19 15.54	N @ 62	2009-09-16 16.42	Y	2009-09-25 16.44	Y			Y- in theory

8.1.4 *rbcL* amino acid sequence analysis

Source	Ref.	AA number	Different AA	Sequence difference	Trace Reading	Decision	Other sequences?	Action?
Kew	perf 921	10	S for G	19 - A	f = trim, r = A	A	Pub G = G, others A = S	None
Kew	perf 921	19	V for E	47 - T	f = A (trimmed), r = T	A	All others A = E	Alter sequence
Kew	perf 921	89	P for R	257 - C, 258 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	perf 921	93	K for E	268 - A for G	f = A/G, r = G	G	8 K, rest E, at EE	Alter sequence
Kew	perf 921	97	Y for F	281 - A for T	f = A/T<25%, r = T	A	pub and 876 = F, rest = Y	None
Kew	perf 921	106	N for D	307 - A for G	f = A/G, r = G	G	pub and 3 = D, rest N	Alter sequence
Kew	perf 921	110	K for E	319 - A	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
Kew	perf 921	159	K for R	468 - A for G	f = A/G, r = G	G	6 = K, rest (inc pub) +R	Alter sequence
Kew	perf 921	160	N for D	469 - A	f = A/G, r = G	Trim	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	perf 932	n/a	Frame Shift	69 - A	f = A, r = gap	Removed		
Kew	perf 932	10	S for G	19 - A	f = trim, r = A	A	Pub G = G, others A = S	None
Kew	perf 932	89	P for R	258 - C, 259 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	perf 932	93	K for E	268 - A	f = A/G, r = G	G	8 K, rest E, at EE	Alter sequence
Kew	perf 932	97	Y for F	282 - A for T	f = A, r = A	A	pub and 876 = F, rest = Y	None
Kew	perf 932	106	N for D	307 - A for G	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	perf 932	110	K for E	319 - A	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
Kew	perf 932	139	Q for R	407 - A for G	f = A/G, r = G	G	6 = Q, 1 = P, rest (inc pub) = R	Alter sequence
Kew	perf 932	159	K for R	467 - A for G	f = A/G, r = G	G	6 = K, others (inc pub) = R	Alter sequence
Kew	perf 932	160	N for D	469 - A for G	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	kou 866	10	S for G	19 - A	f = trim, r = A	A	Pub G = G, others A = S	None
Kew	kou 866	89	P for R	257 - C, 258 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	kou 866	93	K for E	268 - A	f = A/G, r = G	G	8 K, rest E, at EE	Alter sequence
Kew	kou 866	97	Y for F	281 - A for T	f = A, r = A	A	pub and 876 = F, rest = Y	None
Kew	kou 866	106	N for D	307 - A for G	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	kou 866	159	K for R	467 - A for G	f = A/G, r = G	G	6 = K, others (inc pub) = R	Alter sequence
Kew	kou 866	160	N for D	469 - A for G	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	mac 898	10	S for G	19 - A for G	f = none, r = A	A	Pub G = G, others A = S	None
Kew	mac 898	89	P for R	257 - C, 258 - T	f = CT, f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	mac 898	93	K for E	268 - A	f = A/G, f = A/G, r = G	G	8 K, rest E, at EE	Alter sequence
Kew	mac 898	97	Y for F	281 - A for T	f = A, f = A, r = A	A	pub and 876 = F, rest = Y	None
Kew	mac 898	106	N for D	307 - A for G	f = A/G, f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	mac 898	110	K for E	319 - A	f = A/G, f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
Kew	mac 898	139	Q for R	407 - A for G	f = A/G, f = A/G, r = G	G	6 = Q, 1 = P, rest = R	Alter sequence
Kew	mac 898	159	K for R	467 - A for G	f = A/G, f = A/G, r = G	G	6 = K, others (inc pub) = R	Alter sequence
Kew	mac 898	160	N for D	469 - A for G	f = A/G, f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	mac 898	187	K for R	551 - A	f = A/G, f = A/G, r = A (trimmed)	G?	4 = K, rest = R	None

Kew	cal 929	10 S for G	22 - A for G	f = none, r = A	A	Pub G = G, others A = S	None
Kew	cal 929	Frame Shift	72 - A	f = gap, f = A, r = gap	Removed		
Kew	cal 929	89 P for R	261 - C, 262 -	f = CT, f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	cal 929	93 K for E	271 - A	f = A/G, f = A/G, r = G	G	8 K, rest E, at EE	Alter sequence
Kew	cal 929	97 Y for F	285 - A for T	f = A, f = A, r = A	A	pub and 876 = F, rest = Y	None
Kew	cal 929	106 N for D	310 - A for G	f = A/G, r = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	cal 929	110 K for E	322 - A	f = A/G, r = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
Kew	cal 929	139 Q for R	410 - A for G	f = A/G, r = A/G, r = G	G	6 = Q, 1 = P, rest = R	Alter sequence
Kew	cal 929	159 K for R	470 - A for G	f = G/A, r = A/G, r = G	G	6 = K, others (inc pub) = R	Alter sequence
Kew	cal 929	160 N for D	472 - A for G	f = A/G, r = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	cal 929	182 P for A	539 - C	f = C, f = G, r = G	G	All A	Alter sequence
Kew	cal 929	187 K for R	555 - A	f = A/G, f = A/G, r = none	G?	4 = K, rest = R	None
Kew	asc 993	10 S for G	14 - A	f = G trim, r = A	A	Pub G = G, others A = S	None
Kew	asc 993	89 P for R	252 - C, 253 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	asc 993	93 K for E	263 - A	f = A/G, r = G	G	8 K, rest E, at EE	Alter sequence
Kew	asc 993	97 Y for F	276 - A	f = A, r = A	A	pub and 876 = F, rest = Y	None
Kew	asc 993	106 N for D	302 - A	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	asc 993	110 K for E	314 - A	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
Kew	asc 993	160 N for D	464 - A	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	ath 923	10 S for G	18 - A	f = trim, r = A	A	Pub G = G, others A = S	None
Kew	ath 923	35 N for D	93 - A	f = G, r = A/G	G	All others D	Alter sequence
Kew	ath 923	89 P for R	256 - C, 257 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	ath 923	97 Y for F	280 - A	f = A, r = A	A	pub and 876 = F, rest = Y	None
Kew	ath 923	106 N for D	306 - A	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	ath 923	110 K for E	318 - A	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
Kew	ath 923	139 Q for R	406 - A	f = A/G, r = G	G	6 = Q, 1 = P, rest = R	Alter sequence
Kew	ath 923	160 K for N	468 - A	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	perf 876	10 S for G	29 - A	f = G trim, r = A	A	Pub G = G, others A = S	None
Kew	perf 876	89 P for R	268 - C, 269 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	perf 876	106 N for D	318 - A	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	perf 876	160 N for D	480 - A	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	delph 938	10 S for G	21 - A	f = A trim, r = A	A	Pub G = G, others A = S	None
Kew	delph 938	89 P for R	276 - C, 277 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	delph 938	97 Y for F	283 - A	f = A, r = A	A	pub and 876 = F, rest = Y	None
Kew	delph 938	106 N for D	309 - A	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	delph 938	137 N for D	402 - A	f = A/G, r = G	G	2 = N, Rest = D	Alter sequence
Kew	delph 938	139 Q for R	409 - A	f = A/G, r = G	G	6 = Q, 1 = P, rest = R	Alter sequence
Kew	delph 938	160 N for D	471 - A	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence

Kew	and 854	10 S for G	25 - A	f = trim, r = A	A	Pub G = G, others A = S	None
Kew	and 854	89 P for R	280 - C, 281 - T	F = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	and 854	97 Y for F	287 - A	F = A, r = A	A	pub and 876 = F, rest = Y	None
Kew	and 854	106 N for D	313 - A	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	and 854	137 N for D	406 - A	f = A/G, r = G	G	2 = N, Rest = D	Alter sequence
Kew	and 854	160 N for D	475 - A	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	pat 908	10 S for G	22 - A	f = trim, r = A	A	Pub G = G, others A = S	None
Kew	pat 908	89 P for R	260 - C, 261 - T	f = CT, r = CT	Accept P	Pub GC = R, others CT = P	None
Kew	pat 908	97 Y for F	271 - A	f = A/G, r = G	A	pub and 876 = F, rest = Y	None
Kew	pat 908	106 N for D	310 - A	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
Kew	pat 908	139 Q for R	410 - A	f = A/G, r = G	G	6 = Q, 1 = P, rest = R	Alter sequence
Kew	pat 908	160 N for D	472 - A	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
Kew	pat 908	187 K for R	554 - A	f = A/G, r = G trim	G?	4 = K, rest = R	None
NHM	A9	10 S for G	19 - A not G	f = G (trimmed), r = A	A	Pub G = G, others A = S	None
NHM	A9	89 P for R	257 - C for G, 258 - T for C	f + r = C, f + r =T	Accept P	Pub GC = R, others CT = P	None
NHM	A9	97 Y for F	281 - A for T	f = A, T<25%, r = A	A	pub and 876 = F, rest = Y	None
NHM	A9	106 N for D	307 - A not G	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
NHM	A9	110 K for E	319 - A not G	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
NHM	A9	160 N for D	469 - A for G	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
NHM	A9	187 K for R	554 - A		G?	4 = K, rest = R	None
NHM	A10	10 S for G	19 - A not G	f = none, r = A	A	Pub G = G, others A = S	None
NHM	A10	89 P for R	257 - C for G, 258 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	A10	93 K for E	268 - A for G	f = A/G, r = G	G	8 K, rest E, at EE	Alter sequence
NHM	A10	97 Y for F	281 - A for T	f = A, T<25%, r = A	A	pub and 876 = F, rest = Y	None
NHM	A10	106 N for D	307 - A not G	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
NHM	A10	160 N for D	469 - A for G	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
NHM	A12	37 * for L	55 - A not T	f = A/T/G, r = T	T	All L	Alter sequence
NHM	A12	47 A for G	85 - C not G	f = A/T/G, r = G	G	All G	Alter sequence
NHM	A12	66 * for W	142 - A not G	f = A/G, r = G	G	All W	Alter sequence
NHM	A12	89 P for R	211 - C for G, 212 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	A12	97 Y for F	235 - A for T	f = A, r = A/T<25%	A	pub and 876 = F, rest = Y	None
NHM	A12	145 N for T	379 - A not C	f = A/C, r = C	C	All T	Alter sequence
NHM	A12	160 N for D	423 - A for G	f = G, r = A/G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence

NHM	D6	10 S for G	19 - A not G	f = G (trimmed), r = A	A	Pub G = G, others A = S	None
NHM	D6	89 P for R	257 - C for G, 258 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	D6	93 K for E	268 - A for G	f = A/G, r = G	G	8 K, rest E, at EE	Alter sequence
NHM	D6	97 Y for F	281 - A for T	f = A/T<50%, r = A	A	pub and 876 = F, rest = Y	None
NHM	D6	102 P for A	295 - C for G	f = G/C/T, r = G	G	3 = P, rest = A	Alter sequence
NHM	D6	106 N for D	307 - A not G	f = G/A, r = G	G?	pub and 3 = D, rest N	Alter sequence
NHM	D6	119 P for S	346 - C not T	f = T, r = T	T	All S	Alter sequence
NHM	D6	136 Q for E	397 - C not G	f = C/G, r = G	G	All E	Alter sequence
NHM	E5	10 S for G	19 - A not G	f = N, r = A	A	Pub G = G, others A = S	None
NHM	E5	89 P for R	257 - C for G, 258 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	E5	97 Y for F	281 - A for T	f = A/T<25%, r = A	A	pub and 876 = F, rest = Y	None
NHM	E5	106 N for D	307 - A not G	f = G/A, r = G	G?	pub and 3 = D, rest N	Alter sequence
NHM	E5	160 N for D	469 - A for G	f = G/A, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
NHM	E6	10 S for G	24 - A not G	f = G (trimmed), r = A	A	Pub G = G, others A = S	None
NHM	E6	89 P for R	262 - C for G, 263 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	E6	97 Y for F	286 - A for T	f = A/T<25%, r = A	A	pub and 876 = F, rest = Y	None
NHM	E8	2 H for S	1 - C not T, 2 - A not C, 3 - C not A	f = none, r = CAC	Accept	2 H, 2 S, 1P	None
NHM	E8	10 S for G	25 - A not G	f = none, r = A	A	Pub G = G, others A = S	None
NHM	E8	89 P for R	263 - C for G, 264 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	E8	97 Y for F	287 - A for T	f = A/T<02%, r = A	A	pub and 876 = F, rest = Y	None
NHM	E8	106 N for D	313 - A not G	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
NHM	E8	110 K for E	325 - A not G	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
NHM	E8	160 N for D	475 - A for G	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence
NHM	G4	10 S for G	30 - A not G	f = G (trimmed), r = A	A	Pub G = G, others A = S	None
NHM	G4	89 P for R	268 - C for G, 269 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	G4	97 Y for F	292 - A for T	f = A, r = A	A	pub and 876 = F, rest = Y	None
NHM	G4	102 P for A	306 - C for G	f = G/C, r = G	G	3 = P, rest = A	Alter sequence
NHM	G4	106 N for D	318 - A not G	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
NHM	G4	121 A for V	364 - C not T	f = C, r = T	T	All V	Alter sequence
NHM	G4	139 P for R	418 - C not G	f = C, r = G	G	6 = Q, 1 = P, rest = R	Alter sequence
NHM	G8	10 S for G	12 - A not G	f = A (trimmed), r = A	A	Pub G = G, others A = S	None
NHM	G8	89 P for R	250 - C for G, 251 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	G8	97 Y for F	274 - A for T	f = A/T<25%, r = A	A	pub and 876 = F, rest = Y	None
NHM	G8	106 N for D	300 - A not G	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
NHM	G8	110 K for E	312 - A not G	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
NHM	G8	160 N for D	462 - A for G	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence

NHM	G9	10	S for G	21 - A not G	f = N, r = A	A	Pub G = G, others A = S	None
NHM	G9	89	P for R	259 - C for G, 260 - T for C	f + r = C, f + r = T	Accept P	Pub GC = R, others CT = P	None
NHM	G9	97	Y for F	283 - A for T	f = A/T<25%, r = A	A	pub and 876 = F, rest = Y	None
NHM	G9	102	P for A	297 - C for G	f = G/C, r = G	G	3 = P, rest = A	Alter sequence
NHM	G9	106	N for D	309 - A not G	f = A/G, r = G	G?	pub and 3 = D, rest N	Alter sequence
NHM	G9	110	K for E	321 - A not G	f = A/G, r = G	G	pub GAG, others GAA, 10 AAA at EE	Alter sequence
NHM	G9	159	K for R	469 - A for G	f = A/G, r = G	G	6 = K, others (inc pub) = R	Alter sequence
NHM	G9	160	N for D	471 - A for G	f = A/G, r = G	G	5 (inc pub) = D, 1 = K, rest = N	Alter sequence